1966

Predicting the academic achievement of Lehigh freshmen

David A. Riemondy
Lehigh University

PREDICTING THE ACADEMIC ACHIEVEMENT OF LEHIGH FRESHMEN

by
David Anthony Riemondy

A Thesis

Presented to the Graduate Faculty

of Lehigh University

in Candidacy for the Degree of

Master of Science in Industrial Engineering

Lehigh University

1966

This thesis is accepted and approved in partial fulfillment of the requirements for the degree of Master of Science in Industrial Engineering.

February 25, 1966

Sutton Monro
Professor in charge

Arthur F. Gones
Head of the Department

ACKNOWLEDGMENTS

I would like to thank Professor Sutton Monro, who was my principal advisor, for his help and guidance in preparing for and in writing this thesis. I would also like to thank Professors John Adams and William Smith, who were the other members of my committee, for their valuable suggestions and criticisms.

This thesis would not have been possible without the support and confidence of Mr. Samuel Missimer, Director of Admission, who made the data available to me and cheerfully answered all of my questions.

I would also like to thank my patient wife for her help in editing and for carefully typing this thesis.

TABLE OF CONTENTS

Abstract	1
Introduction	3
The Search	9
Analysis of Variables	18
Analysis of Results	32
Analysis of Residuals	41
Correlations	49
Effects of the Sample	54
Summary	58
Appendix 1	66
Appendix 2	78
Interviews	87
Bibliography	88
Vita	90

LIST OF TABLES

Table	Title	
1	Sample Data	67
2	List of Factors	10
3	Coefficients of the Prediction Equations Developed in 1961	72
4	Results of Regression 15 and 1961 Equations	73
5	Correlation Coefficients	76
6	Statistical Summary of Variables	77

LIST OF FIGURES

Figure 1 Distribution of Freshman Cums in Sample 45

LIST OF REGRESSIONS

Regression	
1	19
2	21
3	78
4	78
5	24
6	79
7	79
8	80
9	25
10	28
11	80
12	81
13	82
14	83
15	30
16	83
17	84
18	84
19	85
20	39
21	86
22	86
23A	43
23B	43

PREDICTING THE ACADEMIC ACHIEVEMENT OF LEHIGH FRESHMEN

by David Anthony Riemondy

ABSTRACT

Successful prediction of an individual's future performance is sine qua non in industrial employment and college admissions.

This thesis is the second in a series intended to improve prediction (in particular for admission) from knowledge of the applicant's past and the institution's past. It describes a rudimentary search for a model among sixteen factors related to the applicant's past and representing four previously delimited categories: achievements, aptitudes, interests and family.

Seven of the factors contributed significantly in this study to a predictive model: normalized secondary school rank, SAT verbal score, absolute difference between SAT verbal and mathematical scores, size of secondary school class, number of younger brothers, father's graduation from college, and choice of college for specialization within the university.

The correlations between pairs of variables are calculated and an explanation is offered for those that are significant.

The problems associated with the development of a predictive equation that is to be used as a basis for selection, the inherent limitations imposed by the sample and the difficulty of determining the relative validity of any new predictive model are discussed.

Future investigation into the "Undergraduate Information System" and the problems relating to student evaluation, guidance and selection, possible techniques for this investigation and the use of a computer in the system are recommended.

Introduction

The admissions officers at Lehigh University are quite naturally concerned with selecting from a large group of applicants those candidates who, if admitted, will perform better academically than those applicants who are not offered admission. Thus, there is always a need for better methods of predicting academic achievement among the applicants. There is also a need for such a method to be available for use in directing guidance and counseling to the boys once they are admitted. Therefore, any improvement in the precision of the prediction methods is certainly desirable.

The Admissions Office is presently using a set of four equations to predict the second semester freshman average of each candidate. These equations were developed in 1961 using linear regression on data from the class of 1963. The method for developing these equations was presented in a handbook published by the C.E.E.B. for college admissions officers.[1] The independent variables in the equations are the SAT verbal and math scores and the secondary school class rank.[2] Four different equations were developed to be used depending on whether the applicants

1. Duggan, J. M., Hazlett, P. H., Jr., PREDICTING COLLEGE GRADES, A COMPUTATION WORKBOOK FOR ESTIMATING FRESHMAN GRADE AVERAGES FROM HIGH SCHOOL RECORDS AND COLLEGE BOARD SCORES, College Entrance Examination Board, New York, 1961.
2. The secondary school class rank, i.e., 1st, 2nd,... out of a class of 300, is converted to a standard T score with mean 50 and variance 100 that ranges from 20 to 80 by entering a table provided by the College Entrance Examination Board.

wished to be an engineering, arts, business, or arts-engineering student. All the equations use the same three independent variables although the coefficients are different. The coefficients are listed in Table 3, Appendix 1.

Although the Admissions Office has had some degree of success with these formulae, Mr. Missimer[3] has suggested that I might make an effort to improve the precision of prediction. Since the present equations are only accurate within one half a grade point in 75% of the cases, it appears that a better set of predictors could be discovered. Also, because the variance among the SAT scores and high school ranks of the applicants is quite naturally decreasing, it should be advantageous to include some new factors in the prediction equations.

With this bit of encouragement, I'decided to make the search for some new factors the purpose of my thesis. Much of the preliminary work for my thesis was done by Gordon Bradley in his thesis, Lehigh University Undergraduate Information System.[4] Bradley's thesis "describes the information system the University maintains to gather, use, and store data on the undergraduates and boys who apply for admission to the undergraduate college."[5] My

3. Mr. Samual H. Missimer, Director of Admissions, Lehigh University.
4. Bradley, Gordon H., Lehigh Undergraduate Information System, Thesis, Lehigh University, 1964.
5. Bradley, Gordon H., Lehigh Undergraduate Information System, Thesis, Lehigh University, 1964, pg. 1.

thesis represents an introductory and perhaps unsophisticated attempt to analyze and use some of the available information. Further and more sophisticated analyses of this information should be performed in the future.

There are, of course, some inherent limitations of any analysis of this type. First, the samples of applicants available are not random samples of the whole population of applicants. Therefore, no matter how carefully the independent variables are chosen or how well the prediction equation fits the data from the sample, not much can be done to test the power of the equations in distinguishing between the members of the population. The only thing that can be said about any equation developed from the available data is that it tells us how previous freshmen with particular attributes have performed relative to their peers. This in no way guarantees that students with different backgrounds would not have done better or worse if they had been admitted.

Second, the predicted freshman average, is being used to discriminate between those students who will and those who will not be offered admission, when in fact we are really interested in discriminating on the basis of the actual achieved average. If the predicted average and the actual achieved average are anything less than 100% correlated, we cannot discriminate perfectly.

Thus, a prediction equation based upon background information of boys who have been admitted has at least

two theoretical limitations when used to discriminate among applicants. However, since such an equation is used only as a supplement to the judgment of the Admissions Office and as a basis for guidance of the student, it is reasonable to try to improve the precision of prediction.

Some practical problems now prevent precise estimation. It is hard to choose for the equation the best few variables from among the finitely many for which we have values and the infinitely many for which we do not. E. B. Wilson makes this point very well in his book, AN INTRODUCTION TO SCIENTIFIC RESEARCH, when he discusses the search for causes.

> "The difficult part of the process of seeking the cause of observed phenomena is the construction of hypotheses to be tested. Thus, when it was suspected that the bite of an infected mosquito is the cause of malaria, this hypothesis could be rather easily tested. But the evidence for this cause and effect relationship has presumably always been available, and yet the link was only recently discovered."[6]

The Admissions Office does not make extensive use of the computing facilities available on campus, and does not store its data in a form directly usable by the computer. Therefore, currently useful models must be simple and each new variable considered imposes the additional burden of special data preparation.

Finally, the phenomena I am trying to model is complex.

6. Wilson, E. B., AN INTRODUCTION TO SCIENTIFIC RESEARCH, McGraw-Hill Book Company, Inc., New York, 1952, pg. 32.

According to Shepard,

> "Behavioral scientists, as well as physical scientists, often collect large arrays of data. But, highly developed theoretical models, which have proved so useful in the physical sciences, are still largely lacking in the behavioral sciences. Since less structure can be placed on the data from outside, then, a greater demand is placed on the behavioral scientist to proceed, in a purely inductive way, to discover what structure may already exist in the data themselves. Extracting such latent structure from large arrays of empirical data presents a rather difficult challenge; for in many applications, the number of underlying dimensions of the data is not even known."[7]

In this thesis I am attempting to build a mathematical model to describe a social psychological event, using techniques and skills that I have learned as an industrial engineering student.

This brings me to the question of the relationship of this thesis to industrial engineering. The industrial engineer has traditionally been very much involved in the study of the relationship between a workman's efficiency and the working conditions in the shop. Thus, the problem of developing a mathematical model to describe human behavior is very familiar to the industrial engineering student, and a decent method of searching for significant dependent variables is as desirable in industrial engineering as in admissions work.

7. Shepard, R. N., PROCEDINGS OF THE 1964 SYMPOSIUM ON DIGITAL COMPUTING, "Extracting Latent Structure from Behavioral Data", Holmdel Laboratory, Bell Telephone Laboratories, January 30-31, 1964, pg. 51.

Most universities are presently engaged in some level of a search for models that measure a boy's academic potential. The Department of Measurement and Counseling at Carnegie Institute of Technology has been using the available computer facilities to develop and improve its prediction equations.[8] They are presently using an equation that includes the math and science SAT achievement test scores and the secondary school rank in class converted to a percentile.

Carnegie researchers are using multiple linear regression to develop the coefficients for and evaluate the significance of new variables. This method of searching for significant variables to predict some outcome is very often used by behavioral scientists and will be used in this thesis.

8. This information is the result of a private conversation with F. Jewell of the Psychology Department at Carnegie Institute of Technology.

The Search

There are three major types of decisions that have to be made when one is searching for a model that will describe a boy's academic potential. First, one must decide what bits of information should be admitted as possible sources of independent variables. Second, the form of the model must be chosen (linear, quadratic) and some method must be established for judging the value of the contribution that each variable makes to the model. Third, one must decide how, if at all, the information should be transformed before it is included in the model.

The first phase of the development of a mathematical model is the selection of input information. Gordon Bradley has listed about 200 bits of information about each student that are available in the various campus offices. Of these 200 variables, Bradley lists about 60 that are available to the admissions officers in time for them to be used in making a decision. I limited the search to these 60 factors. One reason for this limitation is that the Admissions Office has selected these 60 factors and has, for years, been using them as a basis for judgment. Thus, it is reasonable to suppose that these 60 factors or some subset of them form a natural equation that results in what a good admissions officer will call his intuition.

However, 60 factors is still quite a few, so to reduce this number further I decided to attempt my factor selection such that I would have a set with at least one variable from each of four of the five classifications of information

proposed by Bradley. That is, I wanted one or more indications of the candidate's achievement, aptitude, interests and background. Bradley's other classification, identity, was not considered useful as a predictor. Table 2 shows the factors that I chose to use to begin my analysis.

Table 2.
Factors Included in Analysis

Classification	Variables	Symbol
Achievement	Class rank	HPOS
	Size of class	HSIZE
Aptitude	SAT verbal score	SATV
	SAT math score	SATM
	Two achievement test scores	ACHA ACHB
Background	Father attend college?	FCOL?
	Mother attend college?	MCOL?
	Father living?	FL?
	Mother living?	ML?
	# of older brothers	#OB
	# of younger brothers	#YB
	# of older sisters	#OS
	# of younger sisters	#YS
	Does the boy need financial aid?*	AID?
Interests	Choice of Curriculum	ENG, BUS, A or AE

* The applicant is required to answer the following question on his initial application form, "Do you plan to apply for aid for your freshman year?". No study was made to determine whether or not the boy actually received aid.

There are, of course, other factors that might be logical additions to this set but were not included because they were either not available at the time of admission or not readily reducible to a usable form. Some of the more obvious factors of this type are reading speed and ability, secondary school activities, whether Lehigh is the applicant's first choice, and the quality of the boy's secondary school.

The four classifications of data were chosen by Bradley to represent four independent groupings in the sense that he expected that there would be more correlation between the bits of information within a group than there would be between bits of data from different groups. Thus, if Bradley's classification is accurate, the set of variables that will form the best model for predicting a boy's academic success should contain at least one variable from each of the four groups of information. The equation may, of course, contain more or less than four independent variables because Bradley's groups are only what he thought to be five jointly inclusive, mutually exclusive classifications. There could possibly be more or less such groups. It is possible by extending the methods of this thesis to test the independence of his groups.

The second type of decision concerns the selection of the form of the model and the method by which the significance of the variables will be tested. It is customary in

problems of this nature to assume a linear model. Thus, the technique of multiple linear regression seems appropriate. With this technique the transformed input information stand for the independent variables, and the achieved average or some transformation of it for the dependent variable. The relative significance of an independent variable is determined by testing the difference between the regression coefficient and zero or by testing whether the addition of the variable to the model causes a significant decrease in the residual sum of squares. Any suspected interactions or non-linear relationships can be written into the regression equation by transforming the independent variables. It should be obvious that such non-linear transformations are essentially changes in the form of the model; however, it is traditional to consider the two seperately. Therefore, to avoid confusion, I will consider the model as linear even though some of the transformations may cause it to be otherwise. The Multiple Linear Regression (MLRP) subroutine developed by General Electric for the GE 225 computer is ideally suited for this type of a search for significant variables.[9] The output from this subroutine includes: a listing of the means and variances of each variable; a listing of the covariances and Pearson product correlation coefficients of all possible pairs of variables; a listing of the coefficients assigned to each variable along with their variance; a listing of the results of a F-test which is performed to check whether

9. GE 225 Multiple Linear Regression Program, General Electric Computer Department, Phoenix, Arizona, June 1962, Lehigh Number D3.503.

or not each coefficient is significant; the residual mean square is calculated and printed along with the correlation coefficient between the predicted and observed values; and finally, the observed and calculated values for each data point are printed along with each residual. A subroutine to allow the programmer to print the data points on tape before calling for the MLRP subroutine is also available.[10] Thus, it was possible to record all the data from a sample on punch cards and then to have the computer make any desired transformations before running the regression analysis.

Having selected the information to be used and the form of the model, there remains the rather difficult question of how this information should be transformed so that it is both usable and, what is more difficult, most effectual in terms of the output. This latter objective is perhaps the most vital point of this thesis. There is, to date, no established method by which one can decide beforehand which transformations should be most effective in terms of maximizing the correlation between the output of the model and the achieved result. Therefore, the choice of transformations is necessarily a trial and error process. Hopefully, the results of this thesis will contain some useful variables that represent new, more effective transformations of the input information but it will also contain a report of transformations that do not work. In both cases, it should be a step forward towards the development of an optimum model.

10. GE 225 WIZ Transformations for Multiple Linear Regressions, Lehigh University Computer Center, Bethlehem, Pennsylvania, 1965, Lehigh Number D3.505.

There is one example that illustrates both the importance of and the obscurity of the types of transformations necessary in this type of analysis.[11] If we look at the subgroup of students whose mothers are widows, they appear perfectly normal in reference to the rest of the students at Lehigh. However, is we consider the group of widows' sons who have no siblings, we find that they all do very poorly.

Now, this is a significant piece of information that can be used to direct guidance toward this type of student once he is admitted. However, the question of how one decides to investigate further into the effect of being a widow's son without siblings once the subgroup of widows' sons has been found to be normal is quite perplexing.

There is also the question of the most effective transformation of the output of the model. There is no way of knowing beforehand what form of output will optimize the correlations between the input and the output variables. There are, of course, some basic relationships that should be considered such as the knowledge that if the input to the model is normally distributed, then it may be advantageous to transform the dependent variable so that it is also normally distributed. Thus, in this problem it is more effective to use the actual grade point average than the percentile rank, since the input variables are for the most part normally distributed.

11. Example provided by Mr. S. Missimer.

One transformation that should be mentioned at this point is the technique used to change the secondary school class rank into a normal deviate. The Polya approximation was considered accurate enough for the purpose of this thesis.[12] The general expression for the area (y) of a normal curve between the mean $-\infty$ and some value x is given by the equation:

$$y = -\tfrac{1}{2}(1-e^{-2x^2/\pi})^{\frac{1}{2}} \qquad \text{(eq. 1)}$$

Solving this equation for x as a function of y we get the following two results:

$$x = (-\pi/2 \ln(1-4y^2)),\ y - .5 \qquad \text{(eq. 2)}$$

$$x = -(-\pi/2 \ln(1-4y^2)),\ y \quad .5$$

Now, if y is given by the equation:

$$y = 1-(HPOS/HSIZE+1)) \qquad HPOS = 1,2,3,...$$
$$HSIZE = 1,2,3,...$$

then equations 2 and 2a will transform the high school rank into a normal deviate with mean 0 and variance 1. This transformation does not result in any loss of degrees of freedom because the data was not used to calculate any constants in the transformation.

There are several reasons for normalizing the secondary school class rank. One reason is that it seems reasonable to suppose that the boy who is first in a class of 400

12. Polya, G., PROCEEDINGS OF THE BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, "Remarks on Computing the Probability Intergral in One and Two Dimensions", University of California Press, Berkeley and Los Angeles, 1949.

should be considered more highly than a boy who is first out of 30. Another reason is that the Admissions Office has had some success with this transformation in the equations presently being used.

Having the multiple linear regression and transformation subroutines available and the variables to be tested selected, all that remains is to collect the data and begin testing. There is, of course, a great deal of data available on campus for a study of this type. I decided to use the class of 1965 as a basis of study because it is the most recent class for which four years of data is available if I should need it. In as much as there were 690 students in that class, I decided to use a random sample of students rather than try to work with the whole group. I chose 100 as a sample size that seemed to be the minimum large enough to estimate the mean square error well. The main practical consideration is that the Registrar does not file information about a student's background in a manner that facilitates the collection of such data for large numbers of students. I formed the sample to be studied by numbering the members of the class according to their alphabetic position and then selecting 100 random numbers. The information for each student was then collected from three documents: the admissions data card which is on file in the Admissions Office, the Application for Admission to the freshman class on file in the Registrar's Office and the Registrar's printout of the cums of the class of 1965, dated October 29, 1964.

The entire sample is given in Appendix 1, Table 1.

The majority of the boys, 70, are engineering students. There are 14 business students and 8 each arts and arts-engineering students. The actual proportion of the students in each of the curriculums is .602, .113, .185, and .100 for the Engineering, Business, Arts and Arts-Engineering curriculums respectively. The reason the sample was not stratified according to curriculum is that Mr. Missimer had expressed a desire for one equation for the whole group. He expressed the opinion that the freshman year did not differ in content very much among the four curriculums. Thus, it seemed reasonable to select the sample at random from the whole class.

Analysis of Variables

The data were collected from the Admissions Office and Registrar's files and punched into 100 data cards. A program was written in Lewiz to make the transformations on the raw data and to write the transformed data on tape for input into the MLRP subroutine. Then, by making minor changes in the transformation program and in the MLRP call cards, I was able to run the regression analysis on any set of variables. What follows is a description of and justification for the sequence of changes that constitited the analysis of variables.

To begin the analysis, I included all of the factors listed in Table 2. Three transformations seemed appropriate at the start; the conversion of the secondary school rank to a normal deviate, the averaging of the achievement scores, and the addition of the absolute difference between the SAT math and SAT verbal scores. The reason for using the converted class rank has already been discussed. The averaging of the achievement scores is a device to permit the common consideration of different achievement scores, since the University does not require every applicant to take the same two achievement tests. The absolute difference between the two SAT scores, it was suspected, would measure a boy's degree of specialization. Note that the absolute difference and the normalized class rank are not linear transformations. The other factors were included in the same form as they are listed in Table 1. The results of the first regression are given on the following page.

REGRESSION NO. 1*

Dependent Variable = FCUM

R**2 (Maximum likelihood) .50815993 00
R**2 (Based on unbiased variances) .42033135 00
Residual Mean Square = .37970139 00
Degrees of Freedom = 84
Residual Sum of Squares = 31.89

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.46721056 00	.17529272 00	.71039111 01	3
AVEACH	.21001566-02	.13641750-02	.23700811 01	0
FCOL?	.25966388 00	.15266036 00	.28931470 01	1
HSIZE	.12305353-02	.49863714-03	.60900226 01	2
SATV	.42296990-02	.17611617-02	.57679364 01	2
DIF	.39547124-02	.18669913-02	.44868878 01	2
#YB	-.20589825 00	.93434557-01	.48561295 01	2
HPOS	-.29878086-02	.18889877-02	.25017694 01	0
#YS	.11710223 00	.11137334 00	.11055231 01	0
#OB	.10686927 00	.13275529 00	.64804033 00	0
#OS	-.79933759-01	.14150152 00	.31910844 00	0
MCOL?	.85632416-01	.14415720 00	.35286102 00	0
SATM	-.83491497-03	.18315217-02	.20780710 00	0
FL?	-.12250755 00	.46116032 00	.70570282-01	0
AID?	.38167383-01	.17291664 00	.48720438-01	0

Constant term = -.23675437 01

* This and all subsequent regression tables are essentially reproductions of the MLRP subroutine printouts. The residual sum of squares and degrees of freedom have been added to some tables. The methods used to calculate the values are given in the description of the program (D3.503) on file in the Computer Center. Briefly, the R**2 value is the correlation between the observed and calculated dependent variables. F(T**2) is a F-statistic equal to the (coefficient/S.D. coefficient)^2. The Conf. LVL's are based on the F(T**2) values and levels of 0, 1, 2, and 3 mean confidence levels of less than 90%, 90% to 95%, 95% to 99% and greater than 99% respectively. The numbers are written in floating point format.

Some of the results of Regression 1 are rather surprising. The first thing to note is the negative coefficient for the SAT math score along with the less than 90% confidence level. This, it will be shown later, is the result of using the absolute difference as a variable. However, it is a bit surprising considering the amount of importance placed on this variable by most laymen. Another unexpected result is the lack of a significant confidence level associated with the average achievement score. My original strategy had called for dropping all variables that were not significant at least to the 90% level; however, since Mr. Missimer had hoped this variable would prove significant, it was included in the next regression.

The second regression included seven independent variables. The residual mean square decreased slightly indicating that no significant variables were included in the first regression that were not included in the second.

REGRESSION NO. 2

Dependent Variable = FCUM
R**2 (Maximum likelihood) .47536219 00
R**2 (Based on unbiased variances) .43544410 00
Residual Mean Square = .36980206 00
Degrees of Freedom = 92
Residual Sum of Squares = 34.02

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.68778149 00	.12378073 00	.30874142 02	3
AVEACH	.13515411-02	.11820535-02	.13073272 01	0
FCOL?	.27023024 00	.13090729 00	.42612816 01	2
HSIZE	.68003533-03	.33287571-03	.41734839 01	2
SATV	.44479439-02	.13477132-02	.10892402 02	3
DIF	.40005195-02	.13447977-02	.88495022 01	3
#YB	-.18492725 00	.87536366-01	.44629783 01	2

Constant Term = -.28657716 01

The third regression was run to check the interaction between the SAT math and the absolute difference. (See Appendix 2). The fact that the SAT math score became significant when the difference was excluded is not surprising. However, this change caused a decrease in the significance of the SAT verbal, the number of younger brothers, and the father college coefficients. As a result of these losses in confidence levels the residual mean square increased. This was convincing that the SAT verbal score and the absolute difference between the SAT math and the SAT verbal scores form a better set of predictors than the SAT math-SAT verbal score set. Therefore, it seems logical that some attempt should be made to improve the effect of the absolute difference.

Algebraically the absolute difference assigns equal weights to those boys who have a high math score and low verbal score and those with a high verbal score and a low math score. Since the average math score (658) is 90 points higher than the average verbal score, the large absolute differences are usually associated with high math scores. So, in order to equalize the weights associated with the two types of differences, 50 was subtracted from the SAT math minus the SAT verbal difference before taking the absolute value. This meant that a boy with a 750 math and 600 verbal score would be weighted the same as a boy with a 650 verbal and a 600 math score. The results of the regression with this new variable substituted for the absolute difference showed a slight increase in the residual mean square. (See Regression 4, Appendix 2.) There was a

drop in the confidence level of the coefficient of the SAT verbal score and the coefficient of the verbal score decreased by almost 50% indicating that the new variable had picked up some of the discriminating power of the SAT verbal score. The net effect, however, was a decrease in the correlation between the observed and the calculated averages and so the set of variables used in the second regression was still preferred.

Another interesting change that should be mentioned is the increase in the confidence level associated with the average achievement score. This seems to indicate an interaction between this variable and either the absolute difference or the SAT math score or both since the latter two have already been shown to interact. A look at the correlation coefficients of these three variables shows very clearly what these interactions are.[13]

VAR.	SATM	SATV	AVE	DIF
SATM	1			
SATV	.27	1		
AVE	.61	.36	1	
DIF	.46	-.55	.19	1

From this table it is obvious that the SAT math score and the average achievement score are highly correlated as are the absolute difference and the SAT math score. Thus, these three variables probably measure the same factor in the student's aptitude. The large negative correlation between the absolute difference and the SAT verbal score is a reflection of the fact that the math score is usually higher than the verbal score and that large differences

13. The values were taken from Table 5 in Appendix 1.

are usually associated with low verbal scores.

REGRESSION NO. 5

Dependent Variable = FCUM

R**2 (Maximum likelihood) .46790705 00
R**2 (Based on unbiased variances) .43357847 00
Residual Mean Square = .37102410 00
Degrees of Freedom = 93
Residual Sum of Squares = 34.51

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.72330157 00	.12001679 00	.36320747 02	3
SATV	.52944052-02	.11280335-02	.22028776 02	3
DIF	.47206719-02	.11901446-02	.15732878 02	3
#YB	-.19177424 00	.87475472-01	.48062674 01	2
HSIZE	.72283627-03	.33131028-03	.47600339 01	2
FCOL?	.28197954 00	.13071881 00	.46532758 01	2

Constant term = -.26828868 01

In the fifth regression the average achievement score was excluded from the set of variables included in the second regression. There was no significant change in the residual mean square. Thus, it can be concluded that the average achievement score does not add anything to the prediction equation that is not encompassed by the verbal score and the absolute difference. The F(T**2) values for both these variables almost doubled when the average achievement score was dropped indicating an increase in the significance of these two variables.

Having eliminated the average achievement score as an independent variable, I thought that the maximum achievement score might provide some additional information. The sixth regression was run to test the significance of this variable

and the results were that the maximum achievement score, like the average achievement score, does not add anything to the precision of the regression equation. (See Appendix 2)

In the next two regressions the number of other siblings in the family variable was tested. The result was that the variable is not statistically significant. (See Regressions 7 and 8, Appendix 2.)

At this point in the analysis I was satisfied that I had found some new variables that would add to the precision of the prediction equation. In order to test this assumption, I ran the ninth regression analysis using the three variables that are presently being used by the Admissions Office.

REGRESSION NO. 9

Dependent variable = FCUM

R**2 (Maximum likelihood) .35352509 00
R**2 (Based on unbiased variances) .33332275 00
Residual Mean Square = .43669479 00
Degrees of Freedom = 96
Residual Sum of Squares = 41.92

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.70258844 00	.12302521 00	.32614728 02	3
SATM	.32868404-02	.10933601-02	.90371361 01	3
SATV	.20229679-02	.10525833-02	.36937289 01	1

Constant Term = -.22091180 01

The effect of using the six variable equation is calculated below.[14]

Source of Variation	df	Sum of Squares	Mean Square	F-Ratio
6 variable regression, No.5	6	30.35		
3 variable regression, No. 9	3	22.93		
Gain due to 6 variable regression	3	7.42	2.47	6.66
6 variable residual	93	34.50	.371	
Total	99	64.85		

The F-ratio at the 99% confidence level for 93 and 3 degrees of freedom is less than 4.04. Therefore, it can be concluded that the six variable equation does make a significant contribution to the variance explained by the regression equation if the choice of curriculum is to be ignored. However, the Admissions Office does include the effect of the choice of curriculum by using a different regression equation on each of the four subgroups. So, it is necessary to compare the residual sum of squares of the new six variable regression equation with the sum of the squared residuals that result when the four three variable equations now being used are applied to the same sample. The resulting sum of squared residuals is 39.04[15] or 4.54 units larger than the residual sum of squares due to the six variable regression. If I assign two degrees of

14. Wert, Neidt, and Ahmann, STATISTICAL METHODS IN EDUCATIONAL AND PSYCHOLOGICAL RESEARCH, Appleton-Century-Crofts, Inc., New York, 1954, Method description - pg 244.

15. From Table 4

freedom to this increase the resulting F-ratio is still a significant 6.12.

Thus, I seem to be able to account for about 10% more of the variation in the dependent variable with a single equation using six independent variables than is accounted for by the four equations now being used by the Admissions Office. It might, however, be possible to improve the equation somewhat by including the curriculum as a variable.

There are several reasons for supposing that the choice of curriculum would have some effect on the freshman cumulative average. First, it is reasonable to expect that the choice of curriculum is a reflection of the boy's interest and background and that there will be less variation in the responses to the academic and social stimulus of Lehigh among the boys who choose the same curriculum than there will be between the boys of different curriculums. Second, there may be a difference in the academic stimuli provided by the different curriculums.

One way to include the effect of the choice of curriculum in the equation is to arbitrarily assign a value of -1 to the arts and arts-engineering students, a 0 to the engineering students and a +1 to the business students and to include this -1, 0, +1 variable in the regression analysis. The reason for this particular choice of values is that, for my sample, the average cums for the groups vary in this manner. The results of including this variable in the regression equation are quite encouraging.

REGRESSION NO. 10

Dependent Variable = FCUM

R**2 (Maximum likelihood) .52043158 00
R**2 (Based on unbiased variances) .48394267 00
Residual Mean Square = .33803396 00
Degrees of Freedom = 92
Residual Sum of Squares = 31.10

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.77763641 00	.11582862 00	.45073515 02	3
SATV	.60663683-02	.11038382-02	.30202768 02	3
DIF	.45883256-02	.11367662-02	.16291692 02	3
CURR*	.35989314 00	.11337674 00	.10076256 02	3
HSIZE	.72984434-03	.31624568-03	.53261241 01	2
FCOL?	.26593365 00	.12487436 00	.45352372 01	2
#YB	-.41235278 00	.84935110-01	.28090364 01	1

Constant Term = -.31785314 01

* CURR (Curriculum): Eng = 0; Bus = +1; A-AE = -1

The residual sum of squares has decreased by 3.3 points to 31.2. Dividing this decrease by the new residual mean square yields an F-ratio of 9.4 indicating that the choice of curriculum is a significant variable that should be included in the set of predictors in some manner. Regressions eleven through fourteen were performed to see if the addition of the curriculum variable caused any of the other variables to become more important. (See Appendix 2)

Another way to handle the effect of the choice of curriculum is to subtract the mean cumulative average of the subgroup from the actual average of each boy, i. e., to subtract 1.94 from each engineering student, 2.14 from each business student, 1.60 from each arts student and 1.96 from each arts-engineering student. Then run the regression analysis using this difference as the dependent variable. This was done in the fifteenth analysis. The independent variables are the same as those used in the fifth regression.

REGRESSION NO. 15

Dependent Variable = FCUM - Mean College Cum.

R**2 (Maximum likelihood) .50751421 00
R**2 (Based on unbiased variances) .47574093 00
Residual Mean Square = .350309571 00
Degrees of Freedom = 89
Residual Sum of Squares = 31.18

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.78450773 00	.11408283 00	.47288322 02	3
SATV	.53957914-02	.10722603-02	.25322689 02	3
DIF	.45369684-02	.11313005-02	.16083315 02	3
HSIZE	.68661594-03	.31492936-03	.47533681 01	2
FCOL?	.25735848 00	.12425570 00	.42898719 01	2
#YB	-.15653036 00	.83150438-01	.35437907 01	1

Constant Term = -.46769986 01

The results of this analysis are very similar to those of the tenth regression. The residual sum of squares is essentially unchanged and there is very little change in the coefficients of the independent variables.

Regressions sixteen through nineteen were run to see if any other variables or combinations of variables would be significant and as before, none were. (See Appendix 2). Thus, there seems to be little advantage to using the transformed dependent variable over using the three valued independent variable to account for the effect of curriculum except that transforming the freshman average separates the arts and the arts-engineering students. Making the transformation also permits a more accurate representation of the differences between the average scores of the four groups although in practice it would probably be just as effective to use the ratio of these averages and to include all four of the ratios in place of the 0, +1 and -1 values used in the fifteenth equation. Whichever technique is used, the resulting regression equation will reduce the residual sum of squares from 38.8 (from Table 6) for the equations now being used to about 31.2. This means that the unexplained variation in the freshman average can be reduced by almost 20% by using the new set of variables.

Analysis of Results

Having selected the variables used in the fifteenth regression analysis as the best set available to be used to predict academic success on the basis of their ability to decrease the residual sum of squares, it might be interesting to see how much each variable contributes to the final regression equation. To do this, I dropped one variable at a time from the final equation and ran the six different six variable regressions. The elimination of any one variable from the regression will, of course, cause an increase in the residual sum of squares. This increase can be interpreted as the contribution made by the eliminated variable to the final regression equation. To establish the relative significance of each variable the F-ratio is formed between the mean square of the gain in the residual sum of squares and the residual mean square of the final equation.[15]

The results of this analysis are tabulated below:

Variables	Residual			Gain			Ratio*[16]
Excluded	df	ss	ms	df	ss	ms	
None	89	31.18	.3503	-	-	-	-
CON HS	90	47.03	.5225	1	15.85	15.85	45.24
SATV	90	39.67	.4408	1	8.49	8.49	24.23
DIF	90	36.57	.4063	1	5.39	5.39	15.39
HSIZE	90	32.77	.3641	1	1.59	1.59	4.54
FCOL?	90	32.61	.3623	1	1.43	1.43	4.25
#YB	90	32.36	.3595	1	1.18	1.18	3.37
Curriculum	93	34.50	.3710	4	3.32	.83	2.38

*F-Ratio 95% confidence level - 80,4df=2.48; 80,1df=3.96.

15. Wert, J.E., Neidt, C.O., Ahmann, J.S., STATISTICAL METHODS IN EDUCATIONAL AND PSYCHOLOGICAL RESEARCH, Appleton-Century-Crofts, Inc., New York, 1954, pgs. 237-249.

16. These values are equivalent to the F(T**2) values in Regression 15 with a slight difference due only to the subtraction of four degrees of freedom from the four curriculum averages used in calculating the dependent variable.

There are four degrees of freedom associated with the curriculum variable because four averages have been calculated from the sample data. The choice of curriculum has a relatively small F value. However, this should not be taken as an indication that the choice of curriculum is not an important factor. Rather, it should be recognized as a reflection of the fact that this method of including the choice of curriculum in the equation is not very effective. The reasons for including the choice of curriculum as a variable have already been stated and nothing has been discovered in the analysis to cause the effect of the choice to be suspected. The lack of significance associated with the number of younger brothers should cause that variable to be viewed with skepticism. In fact, only the first three variables of the table have F-ratios greater than the 99% F ratio of 6.96 and 1 degree of freedom; therefore, these are not subject to suspicion. On the otherhand, the three other background variables should be subjected to further analysis if they are to be considered important as predictors of academic achievement.

Whenever one uses multiple regression to sort out of a group of variables those variables that are most significant in terms of their ability to predict some outcome, he must take care in the final analysis to find some logical explanation for the significance of the particular variables selected. This is especially important when only one set of data has been used to develop the regression equations because it is possible that the resulting significance

is due to some uniqueness of the sample and not the population.[17]

In addition to providing such logical explanations, it is advisable to form at least one other sample, preferably from another time period, and to re-calculate the regression coefficients for the set of independent variables in question. I was not able to obtain a second sample to be included in this thesis; however, I did split the original sample in half and develop a regression equation for both halves. (See Regressions 21 and 22, Appendix 2). In both samples the coefficients for the converted high school standing, the SAT verbal and the absolute difference scores were almost the same as the coefficients from Regression 15. The other three variables exhibited a large disparity both between samples and with the coefficients from Regression 15. This variation in the dichotomous variables could be due to the limiting of the dependent variables with respect to independent variables. The results of this analysis, therefore, tend to lend credibility to the two variables based on the SAT scores and to the secondary school standing. However, the other three variables must be investigated further and justified logically if they are to be considered valid.

17. Ezekiel, M., Fox, K. A., METHODS OF CORRELATION AND REGRESSION ANALYSIS LINEAR AND CURVILINEAR, John Wiley and Sons, Inc., New York, 1959, Third Edition, pg. 297.

The fact that sons of male college graduates do better on the average, as indicated by the positive coefficient of this dichotomous variable in the regression equation, than those boys whose fathers did not attend college is a result that one might expect for various reasons. For example, it is reasonable to suppose that the male college graduate works with paper and considers it important and that a boy who is exposed to such habits and ideas at home is likely to be a better student than one who has not. Also, since a boy's freshman year at Lehigh is usually a new and strange experience socially, it is possible that the son of a college graduate will have been better prepared by his father to handle the problems of finding new friends and adjusting to the new environment.

The positive coefficient associated with the size of the secondary school could be caused by one or both of two factors. First, it could be that the larger secondary schools are giving the boys better academic preparation on the average than the smaller schools. Second, the social experience a boy gains in a larger secondary school could be preparing the boy better for the social shock of being away from home and making new friends.

The fact that the boys from larger secondary schools tend to do better at Lehigh than the boys from the smaller schools is a trend the Admissions Office has noted in past studies. This may seem a bit unusual since most of the private preparatory schools are relatively small (less

than 300 graduates each year.)[18] These past studies have also indicated that there is a higher percentage of failures among the preparatory school graduates than among the public school graduates. It may be possible that the reason for this is that Lehigh is not getting the top students from these schools because these boys are being admitted to the Ivy League schools. It is also possible that those preparatory students who do come to Lehigh do so because they were not accepted by an Ivy League school and are, therefore, disappointed. However, Mr. Missimer tells me that this same tendency for the preparatory school graduates to do relatively poorly was exhibited in studies conducted by the Ivy League schools.

This tendency, if it is statistically significant, causes several questions to become important: What is the advantage to be gained by the boy whose parents have spent extra sums of money to send him to a private preparatory school, if the public schools are better preparing their graduates for college?; Is Lehigh affecting the preparatory school graduate in a manner that is somehow different from the way in which it affects public school graduates?; or, Is the boy who attends a private preparatory school basically different in some respect from the boy who attends public school?. The answers to these questions do not seem to be available at this time.

The other background variable that needs justification is the number of younger brothers. This variable is assigned a negative coefficient by the regression analysis

18. I refer to privately owned and operated high schools, Catholic high schools excluded.

indicating that the boys with no younger brothers do better on the average than the boys with younger brothers. The justification for this is not apparent, but it could be related to the fact that there may be a greater amount of financial pressure on a boy who has one or more younger brothers. Actually, the reason for the significance of the younger brothers, if a causal relationship exists, is most likely psychological and, therefore, beyond the scope of this thesis. However, it does seem reasonable to suspect that the family structure should be related in some way to academic achievement and, therefore, it seems likely that the significance associated with this variable could be due to a causal system and not just to some uniqueness of the sample.

Thus, the two background variables, the number of younger brothers and whether the father attended college, may have some logical justification and bear some relationship to the causal system. Therefore, it is reasonable to include these two variables in the selection equation. The secondary school size is a factor that seems to have had an effect in the past and should be included and investigated further. The SAT verbal score and the converted class rank are widely accepted indicators of academic achievement and the fact that they are so highly significant in this sample can be accepted as an indication of a valid causal system. The only question that remains is what is the best form for including these variables in the model. The absolute difference between the SAT math and the SAT verbal score is a new transformation that is highly

significant. It seems to be a more effective indicator than the traditional SAT math score when it is used in connection with the secondary school rank and the SAT verbal score. Therefore, I would suggest that the normalized secondary school rank, the SAT verbal score, the absolute difference between the SAT verbal and SAT math scores, and the two background variables, number of younger brothers and whether the father is a college graduate, be used to develop a new set of prediction equations.

I have suggested that the choice of curriculum is an important variable and that it should not be included in the regression equation, but that a different regression equation should be developed for each of the four curriculums. Essentially, what this means is that I think that the choice of curriculum makes more than a linear contribution to the model. One way to check this hypothesis is to take a different sample of students from each of the four curriculums and to form a regression equation using the same variables for each group and then to compare the coefficients. If the curriculum makes a linear contribution the only significant difference among the equations should be among the constant terms. I did not obtain a large enough sample to provide enough data points in each of the curricula for an analysis of this type. There are, however, a sufficient number of students in the engineering subsample to form a regression equation for this group. The results of this regression are given on the following page.

REGRESSION NO. 20
Engineers Only*

Dependent Variable = FCUM

R**2 (Maximum likelihood) .46998760 00
R**2 (Based on unbiased variances) .46998760 00
Residual Mean Square = .36134885 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.80000546 00	.15865961 00	.25424543 02	3
SATV	.52604413-02	.14601688-02	.12978910 02	3
DIF	.37324291-02	.15144487-02	.60739888 01	2
FCOL?	.23266830 00	.14843144 00	.24570990 01	0
HSIZE	.51769074-03	.42577083-03	.14783899 01	0
#YB	-.11331584 00	.10534417 00	.11570716 01	0

Constant Term = -.26990245 01

* 70 Data points.

The interesting thing about this regression is that the three background variables and the absolute difference have lost significance. This leads me to believe that the significance of these variables is due, in part, to their relationship to the choice of curriculum. If this is true, then it may be possible to develop a single equation that will serve as a more valid model for predicting academic success than can be developed by considering each curriculum separately.

Analysis of Residuals

Before one can conclude that a regression analysis includes all the significant information contained in the set of available data and uses the most effective transformations of that information (model), there must be some analysis of the residuals (errors) that result when the predicted outcome is compared with the actual outcome. Also, the hypothesis that the sample of students forms a homogeneous group in the sense that all the boys respond in the same general way can be tested by an analysis of the residuals.

The problem of determining whether or not there has been optimal utilization of the available information is, of course, quite difficult since it is obviously dependent on the transformation of the information. However, one test that is advisable in helping to decide whether or not any usable information has been excluded is to form what seem to be the most logical variables from the information that is not included in the final regression equation and to run a regression analysis using these variables as the independent variables and the residuals from the final regression equation as the dependent variables. If all the useful information has been included in the final equation then there should not be any correlation between the unused variables and the residuals. If, however, one of the excluded bits of information is important, then there will be some correlation between that variable and the residual if the information has been expressed in

the proper form. Regression 23A on page 43 gives the results of an analysis of this nature using the nine independent variables that were dropped from the original set. Note that the residual mean square for this regression of .3326 is not much smaller than that of the fifteenth regression analysis. Therefore, there does not seem to be any information contained in the excluded variables that is not included in the variables used in the fifteenth regression analysis.

REGRESSION NO. 23A

Dependent Variable = Residual from Regression 15

R**2 (Maximum likelihood) .39806653-01
R**2 (Based on unbiased variances) -.56212680-01
Residual Mean Square = .33262698 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
AVEACH	.13309474-02	.11504602-02	.13383774 01	0
HPOS	-.82096609-03	.10310087-02	.63405333 00	0
#YS	.56005843-01	.98426811-01	.32377241 00	0
#OB	.10828420 00	.12282358 00	.77726073 00	0
MCOL?	.10286006 00	.12988785 00	.62712847 00	0
SATM	-.89261878-03	.12103804-02	.54386142 00	0
#OS	-.68754445-01	.12518658 00	.30163796 00	0
FL?	-.32920245-01	.42047490 00	.61298325-02	0
AID?	.33292770-02	.15681386 00	.45074504-03	0

Constant Term = -.19888004 00

REGRESSION NO. 23B

Dependent Variable = Residual from Regression 15

R**2 (Maximum likelihood) .51082526 00
R**2 (Based on unbiased variances) .50583368 00
Residual Mean Square = .15562495 00

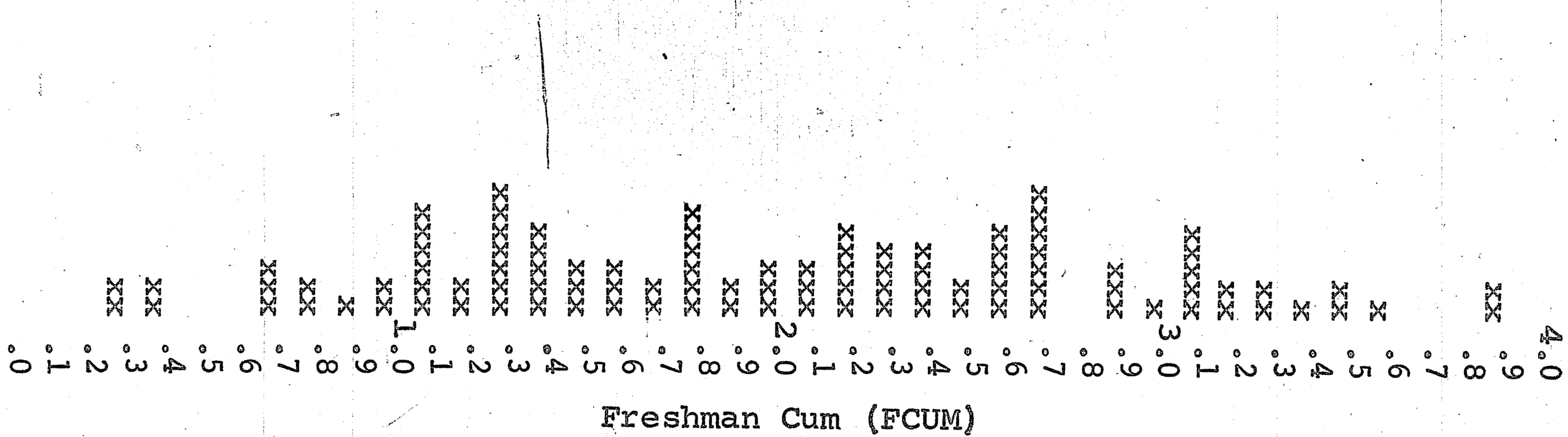
Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
FCUM	.49557346 00	.48988127-01	.10233741 03	3

Constant Term = -.94119298 00

A useful technique for checking the form of the model is to plot the residuals against the observed values and to fit a regression line to these points. If the form of the model is correct, i.e., if the input and output information has been transformed correctly, then there should not be any correlation between the dependent variable and the residuals. Regression 23B(page 43) was run using the freshman cum as the independent variable and the residual from the fifteenth regression as the dependent variable. Note there is a high degree of positive correlation (.70) between these two variables indicating that equation fifteen tends to over-estimate the performance of the poor students and to under-estimate the performance of the good students. This indicates that I have probably failed to represent the variables in the best possible form. This could be due to the fact that the dependent variable is bounded. It might be more advantageous to rank the students and then normalize the ranks so that the dependent variable will be closer to being truly normal. This type of correlation between the error and the dependent variable will result if the dependent variable is uniformly distributed. The distribution of freshmen cums observed in this sample is neither normal nor uniform but the frequency of observed cums around the plus or minus one standard deviation from the mean is quite large. The freshmen cums are plotted in Figure 1 on the following page.

Figure 1

Distribution of Freshmen Cums in Sample



The histogram almost looks bimodal. This could cause the slope of the regression equation to be controlled by the points at plus or minus one standard deviation. I suggest, therefore, that in any further attempt to increase the validity of the prediction equation the normalized ranks be used as the dependent variable in the regression equation. This may cause some problems in the interpretation of the prediction at first; however, the admissions officers and the other users of these predictions should find that the normalized rank is more helpful provided they take advantage of the available computer facilities to make the conversions.

The question about the homogeneity of the sample can be answered by plotting the residuals on probability paper. If the residuals are normally distributed with mean zero and variance equal to the residual mean square, it can be concluded that the performance of all the boys responded in the same general way to the independent variables. Plotting the residuals on probability paper, therefore, should result in a straight line. If any of the points deviate from the line by very much then these boys should be excluded from the sample used to develop the regression equation and an attempt should be made to determine why they are different from the rest of the sample. When the residuals from the fifteenth regression are plotted, all of the points fall very near to a straight line. Therefore, I can not reject the hypothesis that the sample is a homogeneous group from this point of view.

However, if the residuals are analysed by curriculum (See Table 4, Appendix 1) it turns out that 12 out of the 14 business student's residuals are positive. This result has a probability of .013 of occurring by chance alone if a simple sign test is used.[20] This result should add strength to the suggestion that the prediction equation be developed separately for each college, especially in light of the fact that the dependent variable used in the fifteenth regression equation had already been corrected for differences in the mean cum for each of the four curriculum choices.

Thus, the analysis of the residuals has disclosed four rather useful pieces of information: (1) It is possible to say that if the discarded information has been expressed in the most useful form then there is nothing available from that information that is not included in the variables used in the fifteenth regression analysis;

20. Probability of 12 or more the same out of 14 =

$$2\left(\frac{14!}{12!\ 2!}\left(\tfrac{1}{2}\right)^{14} + \frac{14!}{13!\ 1!}\left(\tfrac{1}{2}\right)^{14} + \frac{14!}{14!\ 0!}\left(\tfrac{1}{2}\right)^{14}\right) =$$

$$\tfrac{1}{2}^{13}(91 + 14 + 1) = \frac{106}{8192} = .013$$

(2) It appears that there may be some better method of transforming the information used to improve the validity of the regression equation; (3) It seems that all of the sample forms a homogeneous group, but; (4) the business students as a group seem to differ from the rest of the sample.

Correlations

One of the more useful by-products of the linear regression subroutine is a listing of the correlation coefficients between all the possible pairs of variables included in the analysis. Some of these correlations have already been discussed where they were used to make decisions about the choice of model. However, there are some other relationships that should be pointed out here since they will probably be useful to anyone who is interested in pursuing this problem further. All of the coefficients are given in Table 5 in Appendix 1. They are all calculated using the Pearson product-moment method.[19] Consequently, the correlations between any two dichotomous variables may be misleading. However, it should be possible to pick out pairs of variables that seem to be more highly correlated than others, and, if one is interested, to investigate the relationships more carefully using contingency tables or by calculating the tetrachoric correlation coefficients.[20]

19. Wert, J. E., Neidt, C. O., Ahmann, J. S., STATISTICAL METHODS IN EDUCATIONAL AND PSYCHOLOGICAL RESEARCH, Appleton-Century-Crofts, Inc., New York, 1954, pg. 78.
20. Ferguson, G. A., STATISTICAL ANALYSIS IN PSYCHOLOGY AND EDUCATION, McGraw-Hill Book Co., Inc., New York, 1959, chapters 11 and 13.

To test the significance of a correlation coefficient a t-test[21] can be used where:

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

N = Number in sample
N-2 = Number degrees of freedom
r = Correlation coefficient

The t value for a 99% confidence level two-tailed test is about 2.603, so for a coefficient to be significant at this level it must be greater than .255 or less than -.255. At 95% the t value is about 1.973 and the coefficient must be greater than .195 or less than -.195. Note that this test of significance is strictly applicable to Pearson product-moment correlation coefficients only when the two variables are normally distributed. If they are not normally distributed then the correlation coefficients could be over estimated. Therefore, this test should only be used to provide minimum values with which the correlations listed in Table 5 can be compared.

Some of the more interesting of the correlations are those associated with the whether-the-father-went-to-college (FCOL?) variable. Notice that the correlation between the FCOL? variable and the normalized secondary school standing (CON HS) is -.28. This means that the sons of college graduate fathers tend to do porrer on the average in secondary school than the boys whose fathers are not college graduates. This is in contrast to the effect of the FCOL? variable in the regression equations where the

21. Ferguson, G. A., STATISTICAL ANALYSIS IN PSYCHOLOGY AND EDUCATION, McGraw-Hill Book Co., Inc., New York, 1959, pg.152.

sons of college graduates are predicted to do better on the average. Another interesting correlation is the -.32 coefficient between the FCOL? variable and the engineering; non-engineering dichotomous variable. This indicates that the boys whose fathers are not college graduates tend to choose engineering at Lehigh. Since both of these variables are dichotomous, I set up the following two by two contingency table:

		Engineering		
		Yes	No	
FCOL?	Yes	36	25	61
	No	34	5	39
		70	30	100

The resulting chi square is 10.64 with one degree of freedom. This is greater than the 99% confidence level chi square of 6.64, so the hypothesis that there is no correlation should be rejected.

Some other large coefficients involving the FCOL? variable that might bear further investigation are the +.37 between FCOL? and MCOL?, and the -.17 between FCOL? and AID?.

The .31 correlation coefficient between the number of older brothers (#OB) and the boys who express a need for financial aid (AID?) on their initial application is expected as are the negative correlations between the AID? and the MCOL? and FCOL? variables, and the .23 correlation between the AID? and the number of siblings (#SIBS) variables. These indicate that the boys who come from smaller families, who are sons of college graduates, and who have no older brothers tend not to seek financial aid.

The correlations between the number of siblings and the number of younger brothers and the number of younger sisters indicates that most of the boys in this sample are older than their siblings. I doubt whether this will be true of all the classes attending Lehigh. I suspect that it is partly due to the fact that the majority of the boys in the class of 1965 were born in 1943 or 1944. It may be worth investigating the family structure further because the number of younger brothers is a variable that seems to be useful in predicting academic success. Perhaps the value of this variable will change as the family structures of the students who enter Lehigh change. Note that for this sample 79% or more of the boys are the oldest boys in their families.[22]

Another interesting coefficient is the .38 correlation between the converted secondary school standing and the engineers, non-engineers dichotomous variable. This means that boys who are high in their class in secondary school tend to want to be engineers and/or that in order for a boy to be admitted to the engineering college at Lehigh he must rank higher in his graduating class than if he wants to enter the arts or business colleges. This correlation along with the negative correlations between the FCOL? and CON HS variables and the FCOL? and Eng? variables might help explain why the FCOL? variable is assigned a positive coefficient by the regression analysis even though there does not seem to be any correlation between the FCOL? and the freshman cum.

22. From Table 6 the average number of older brothers is 21.

It should be noted that all of the correlation coefficients listed by the MLRP Subroutine are zero-order correlations, whereas the regression coefficients are related to the higher order partial correlations as well. The fact that the FCOL? variable is positive and significant indicates that there is some correlation between the freshman cum and FCOL? if some or all of the other variables are held constant.[23]

There are, of course, other relationships between the variables that could be analysed by anyone wishing to continue the search for significant predictors. Only a few of the more apparent and interesting correlations have been mentioned here. Actually, the analysis of the higher order correlations should lead to the selection of the same variables as the regression analysis. The only advantage to using the regression technique is that it is computationally easier. The analysis of partial correlations will provide a great deal more information about the relationships between the variables and may be desirable for that reason.

23. Wert, J. E., Neidt, C. O., Ahmann, J. S., STATISTICAL METHODS IN EDUCATIONAL AND PSYCHOLOGICAL RESEARCH, Appleton-Century-Crofts, Inc., New York, 1954, Chapter 13.

Effects of the Sample

It should be obvious that the results of this type of regression analysis are very much dependent on the sample of students on which the analysis is performed. Of course, these effects are less reparable if the resulting equation is to be used as a basis for screening applicants than they are when the equation is to be used to direct guidance once the boy has been admitted.

Some of the problems involved with using this type of analysis to develop an equation for screening applicants have already been mentioned in the Introduction. However, they will be re-emphasized here because they must be understood before any attempt is made to use the results of this thesis as a basis for the development of better prediction equations. The crux of the problem associated with the sample used is that it represents a subsample of the population of applicants who have already been screened by the Admissions Office and found to be acceptable according to a group of criteria one of which was a predicted average. Thus, it is likely that the distribution of the predicted averages, based on the old equations, of the admitted group will tend to be skewed to the right, even though there is no minimum cut-off for predicted averages. Since the predicted average has in the past been a function of the SAT math and SAT verbal scores, and the secondary school standing, the distribution of these three independent variables may also be skewed. The range associated with these variables and with the predicted average of the sample could be less than the range associated with these factors in the whole population of applicants.

Conversely, the distribution of the achieved averages for the sample has the same range of values as would be achieved by the population. Consequently, any new equation that is developed to arrive at a predicted average, if it contains new variables that are correlated with the variables used in the screening equation but not subject to that screening, will probably result in a predicted average that is more highly correlated with the achieved average than the old equation. This is because the range on the new equation will not be restricted by the screening process. The correlations between the new independent variables and the achieved average will likewise probably be greater than those between the achieved average and the screened variables with which the new variables are correlated. These correlations between the independent variables and observed averages can be corrected to account for this screening if a minimum cut-off point has been established for each of the independent variables and if the other criteria on which the selection is performed are well defined. If, however, there are no minimum cut-off points and if some boys with very low predicted averages have been admitted on the basis of other factors, then there is no way of correcting the correlation for the effect of the screening.

Thus, it is impossible on the basis of any sample that is available at this time to say with a great deal of assurance that the set of variables suggested by this thesis

24. Gulliksen, H., THEORY OF MENTAL TESTS, John Wiley and Sons, Inc., New York, 1950, Chapters 11, 12 and 13.

could indeed yield a predicted average that will be more correlated with the achieved than the equations now in use. The only way the validity of a new equation can be compared with the old equations is to use the new equation to screen an incoming class and see how well the new predictions correlate with the performance. It will not suffice to try the new set of variables on another sample of students that have already been screened, although this additional test might show a "significance-change" in some of the variables and is, therefore, advisable.

The fact that the relative validity of the new regression equation is not readily available from an analysis of this data does not mean that the regression coefficients will necessarily be in error. To the contrary, since the independent variables have not been subject to minimum cut-off points, the regression hyperplane is not restricted. The only problem is that it is impossible to say, with any degree of certainty whether or not the new variables do indeed form a better set of predictors than the old variables.

The accuracy of the regression coefficients are, however, very much dependent on the randomness of the sample with respect to both the dependent and the independent variables. The sample must be selected so that the range of the dependent variable is not restricted with respect to any one of the independent variables unless this restriction is caused by the independent variable.[25]

25. Ezekiel, M., and Fox, K. A., METHODS OF CORRELATION AND REGRESSION ANALYSIS LINEAR AND CURVILINEAR, John Wiley and Sons, Inc., New York, 1959, Third Edition, Chapter 18.

For example, in the sample used in this thesis, the range of the dependent variable for the arts-engineering students is restricted to values between .85 and 1.97. Since the dependent variable is dichotomous, the regression equation will force the regression hyperplane through the average grade point for this range when the boy is an arts-engineering student. If this average is not the average of the population the coefficient associated with the dichotomous variable will be in error. This same sort of error will occur if, for example, the range of the grade point averages is restricted with respect to the SAT verbal scores included in the sample. Therefore, it is advisable, either to check the ranges of the dependent variable associated with each independent variable for each student in the sample or to include as many students as possible in each sample to be used for the development of the regression equations.

Summary

The development of a better model for describing a boy's college potential is desirable. Sophisticated techniques available for developing such a model are not feasible at this time, but a preliminary analysis is needed so that basic information will be available to future investigators. Specifically, it is necessary at this time to attempt to reduce the large amount of information that is available about the student to a small set of variables that seems to contain all the information that is useful for prediction.

This thesis is concerned with extracting significant variables from the information that is available when the boy applies for admission. The techniques used to search for the significant factors were largely intuitive. The method used to check the significance of each variable and to compare sets of variables (models) was multiple linear regression. The residual mean square was used as an indicator of the validity of the model and, the changes in the residual sum of squares as a means of comparing the contributions to the model of the different variables.

The results of this procedure suggest that seven variables are significant indicators of academic success at Lehigh. These variables are: the normalized secondary school class rank, the SAT verbal score, the absolute difference between the SAT verbal score and the SAT math score, the size of the boy's class in secondary school, the number of younger brothers the boy has, whether the

boy's father is a college graduate, and the curriculum that the boy chooses. These seven variables are based on information about the boy's achievements, aptitudes, interests and family background. That is, there is at least one variable from each of these four categories of information as they are defined in Bradley's thesis.

The analysis of these seven variables suggests that the first three of these variables are much more significant than the other four. There is some question as to the best method for including the choice of curriculum in the model. I have suggested that a separate equation be developed for each curriculum. However, an attempt to use the remaining six variables to form a regression equation for the 70 engineering students in my sample indicates that it may be possible that some of the background variables are related to the choice of curriculum. An analysis of the correlation coefficients, for example, shows that there is a tendency for the boys whose fathers are not college graduates to choose engineering. Therefore, it may be possible to include the choice of curriculum in a single model to represent the effect of the curriculum on the student and to include background variables that will measure the student's interest. Note that the fact that the choice of curriculum is included in the model suggests that it is possible for the University to say something to the applicant about his relative potential with respect to the curriculum he chooses.

There is also some question about the significance of the number of younger brothers. I suspect that the

significance of this variable may be unique to the class of 1965 and that it will not be useful for the class of 1970. I have suggested that the structure of the family is probably important but further studies will be needed to find a variable that represents the significant information about the family structure.

There is a great deal of research to be done in the future if the University is going to maximize the accuracy with which it can evaluate the academic potential of applicants and students. In order for this research to be carried forward, it will first be necessary to record the large amounts of data that is available on campus on computer input cards so that the computer can be used to make the analyses.

Finally, I suggest that the University update the present prediction equations and that the variables found to be significant in this thesis be included. I also recommend that the prediction equations be recalculated annually and that a continuing effort be made to improve them.

Suggestions for Future Work

The problem of reducing the large amount of information that the University collects and stores about each student of and applicant to Lehigh is quite broad and complex. This thesis is a preliminary search for the solution to one aspect of this problem, namely the reduction of the large amount of information that is available about the applicant to a form that can be used to predict academic achievement. Because this thesis is an introductory look at the problem of predicting academic achievement it has probably raised more questions than it has answered. There is a great deal of work left to be done in this area if the University is to optimize the benefits to both the institution and the students obtainable through a well organized, scientific study of the "undergraduate information system". What courses this study will take are not obvious, but some specific steps do seem to be logical extentions of what has been done here.

The next step should be a more extensive investigation of the possible factors to be used as a basis for prediction and of the form in which these factors should be represented in the model. This investigation should make corrections for the screening process and include tests of the model. Some variables that were not investigated in this thesis but might be of value are the results of reading tests for both speed and comprehension, the father's occupation, the average number of hours spent in extracurricular activities

each week while in secondary school, the slope of the line connecting the boy's secondary school grade point averages when they are plotted against time, and the way the boy ranks Lehigh relative to the other schools to which he applies. The transformations of the variables should also be investigated further. The effect of the secondary school class rank might be improved if the rank is converted to a standard measure with respect to a recent population of applicants, say the past two years. Likewise, it may help to standardize the SAT scores to the Lehigh norm. However, all of these investigations will be prohibitively laborious unless the present means the University uses to handle and store information are replaced by a well designed computer-oriented information system.

The development of the computer-oriented data processing system is an event that must occur if Lehigh is going to optimize the benefits of making full use of the student information that is available. For example, it should be relatively easy to develop a system for collecting and storing the information received from each applicant on cards and eventually on magnetic tape or disks. Once this is accomplished, the soft wear can be developed to transform and organize this data into any form desired. A technique that is powerful, but requires a large amount of information readily available for input into rather complex computer programs, is Factor Analysis.[26] Shepard's paper, "Extract-

26. Shepard, R. N., PROCEDINGS OF THE 1964 SYMPOSIUM ON DIGITAL COMPUTING, "Extracting Latent Structures from Behavioral Data", Holmdel Laboratory, Bell Telephone Laboratories, January 30-31, 1964.

ing Latent Structure from Behavioral Data", gives a good example of the application of this technique. Some study will be required before this technique can be applied to the problem of predicting academic success.

Another method of solving the problem of prediction and the broader selection problem as well is to program the computer to "learn". That is, to develop the soft wear so that the computer will adjust the prediction model to account for the errors it has experienced in the past. The prediction model is developed and programmed into the system. The variables are input to the system and a prediction is calculated by the computer. The results are observed and re-inputted into the system. The computer then compares the results with the predictions that it made and adjusts the prediction model. The steel industry is presently experimenting with such a technique to "teach" a computer to make decisions about the quantities of alloys, temperatures and melting times necessary to obtain metals with customer specified physical properties. Some work is necessary before Lehigh can use this technique. Criteria for judging the results would have to be established. A statistical test for evaluating the power of a discriminator will have to be selected or developed. Second, it will be necessary for the admissions officers to find out what happens to some of those boys who are not offered admission and to some of those who are but do not accept the offer. This data must then be converted into a form that is compatible with the information collected from boys that attend Lehigh. Finally, and this is true even for the present

method of discrimination, some technique should be established for including time in the model.

The effect of time on the prediction equations has not been considered in this thesis except to say that the equations should be updated each year. The problem is that the group of applicants that applies to Lehigh one year may not be the same on the average as the group that applies the next year. This may be caused by changes in the curricula of the secondary schools, in the financial aid policies of the University or other institutions, in the recruiting practices or in the grading practices of the University, or in a number of other factors. Therefore, it will probably be necessary to update the equation to account for these changes. It may, however, be possible to develop a single equation or set of equations that include the effects of time by using a sample from several time periods. There are two ways by which the effect of time can be established; one is to include the time variable in the equation and to perform the regression analysis on all the data from several time periods at once; the other method is to run the regression analysis on each time period separately and then to plot the coefficients of the independent variables as functions of time and to find the equations that express these coefficients as a function of time. These two methods are not equivalent, and the first is technically preferable.

The development of a model to predict the success of the student beyond his freshman year is another aspect of the problem that needs to be considered because of its value in counseling the student. It could also serve as a basis for measuring the effect of Lehigh upon the student.

One final point that must be the subject of any further considerations of the development or improvement of mathematical models to predict academic achievement is the way in which these models and the experience and knowledge gained through the development and use of these models will effect the character of the University. As the ability to predict academic achievement increases, the amount of information that the University is able to store and interpret about the student will also increase. With this increase in understanding will come an increase in power and, therefore, responsibility. To whom this power will be entrusted and how it will be used to benefit the student are questions that must be answered before, not after, that power is available.

APPENDIX 1

KEY TO TABLE 1

Column	Heading
1	Number assigned to student
2	Student's cumulative grade point average-Spring '62
3	Student's cumulative grade point average-Fall '61
4	SAT verbal aptitude test score
5	SAT math aptitude test score
6	SAT achievement test score
7	SAT achievement test score
8	Rank in secondary school graduating class
9	Size of secondary school graduating class
10	Normalized rank in secondary school graduating class
11	Is the student's father living?
12	Is the student's mother living?
13	Does the student say he needs financial aid?
14	Number of older brothers
15	Number of older sisters
16	Number of younger brothers
17	Number of younger sisters
18	Is the student's mother a college graduate?
19	Is the student's father a college graduate?

* In all yes or no factors, yes is 1 and no is 0.

APPENDIX 1

TABLE 1 - Sample Data

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Engineers																		
1	2.03	1.94	637	687	662	483	8	82	1.30	1	1	0	0	1	0	0	1	1
2	2.91	3.22	510	713	662	620	1	104	2.35	1	1	0	0	0	1	2	1	1
3	2.64	3.06	643	767	800	729	38	313	1.35	1	1	1	0	0	0	1	0	0
4	2.41	2.33	648	800	778	697	9	254	1.81	1	1	0	1	0	0	0	0	0
5	2.06	2.38	563	626	568	652	36	264	1.10	1	1	0	1	0	1	0	1	1
6	.82	.72	540	667	721	522	12	79	1.10	1	1	0	0	1	0	0	0	0
7	2.79	2.65	448	667	701	610	1	34	1.89	1	1	0	0	2	0	0	1	1
8	2.51	2.53	583	697	643	491	7	207	1.83	0	1	0	0	0	1	1	1	1
9	1.33	1.63	515	587	493	505	75	473	1.00	1	1	0	0	0	1	0	1	1
10	.76	1.00	556	678	596	464	11	45	.71	1	1	0	0	1	0	0	1	1
11	2.33	2.38	519	719	614	655	3	237	2.24	1	1	0	0	0	0	0	0	1
12	2.06	2.67	603	687	594	620	46	264	1.45	1	1	1	0	0	1	1	1	1
13	2.79	3.00	475	677	596	505	11	113	1.30	1	1	0	0	0	0	0	0	1
14	2.24	2.24	444	626	707	585	16	406	1.76	1	1	0	0	0	1	0	0	0
15	1.97	2.19	582	652	612	652	7	118	1.57	1	1	0	0	0	0	1	0	1
16	.60	.61	482	707	634	566	8	19	.25	1	1	1	0	0	2	1	1	1
17	1.94	2.25	622	617	493	610	10	97	1.27	1	1	0	0	0	0	0	0	0
18	2.14	1.24	616	714	701	505	148	543	.61	1	1	0	0	0	0	1	0	1
19	2.36	2.47	616	617	568	572	17	266	1.52	1	1	0	0	0	1	1	1	0
20	2.42	2.68	556	757	745	615	17	170	1.29	1	1	1	0	0	0	1	1	1
21	.25	.25	516	609	540	424	27	149	.92	1	1	1	3	1	0	0	0	0
22	.94	1.33	596	696	615	565	66	192	.41	1	1	0	0	0	1	0	0	0
23	2.15	2.56	489	625	587	560	45	627	1.46	1	1	0	0	1	0	0	0	0
24	1.00	1.31	623	607	521	572	42	254	.98	1	1	0	0	0	1	0	0	0

APPENDIX 1

TABLE 1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Engineers (con't)																		
25	1.09	1.00	516	731	502	489	18	66	.62	1	1	0	0	0	1	0	0	0
26	1.57	1.74	547	667	624	622	14	91	1.03	1	1	0	1	1	1	0	0	1
27	1.31	1.22	623	617	656	586	12	160	1.44	1	1	0	0	0	0	0	0	1
28	1.67	2.06	477	565	559	522	23	805	1.90	1	1	0	0	0	0	0	0	1
29	2.56	2.24	489	637	445	631	40	354	1.21	1	1	0	0	0	1	0	0	0
30	1.21	1.75	516	635	464	500	50	277	.92	1	1	0	0	0	1	0	0	1
31	2.79	3.12	610	757	778	609	41	277	1.05	1	1	0	0	0	0	1	1	1
32	1.76	1.94	563	577	455	522	14	200	1.48	1	1	0	0	0	0	0	1	0
33	3.31	3.47	650	757	734	654	34	543	1.53	1	1	1	0	0	1	0	0	0
34	3.19	3.41	589	717	653	591	50	438	1.20	1	1	0	0	0	1	0	1	1
35	.67	.75	543	617	577	522	119	345	.40	1	1	0	0	0	2	2	0	1
36	.75	.38	497	591	426	542	20	175	1.21	1	1	1	0	0	2	0	0	0
37	1.12	1.00	547	596	549	565	60	402	1.04	1	1	0	0	0	1	0	0	0
38	1.17	1.00	510	733	634	516	73	805	1.34	1	1	0	0	0	1	0	0	0
39	2.03	2.44	468	637	582	552	10	401	1.96	1	1	0	1	0	0	0	0	0
40	2.21	2.63	510	696	615	576	27	320	1.38	0	1	1	0	0	1	0	0	1
41	1.18	1.13	475	697	587	691	13	354	1.79	1	1	0	0	0	0	0	0	0
42	1.28	1.19	522	495	530	528	21	135	1.02	1	1	0	0	0	1	0	0	0
43	1.15	1.56	509	661	559	462	48	407	1.19	1	1	0	0	0	3	0	0	1
44	2.32	2.67	657	747	712	628	12	207	1.57	1	1	1	2	0	2	2	0	0
45	2.68	2.63	589	625	540	587	19	202	1.32	1	1	0	0	1	0	1	1	0
46	.60	.60	603	547	502	522	6	82	1.46	1	1	1	0	0	0	0	0	0
47	3.91	3.83	657	717	662	697	11	368	1.88	1	1	1	1	0	0	0	0	0
48	1.15	1.38	482	598	615	660	5	115	1.72	1	1	0	0	0	0	1	0	0

APPENDIX 1

TABLE 1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Engineers (con't)																		
49	.53	.35	549	622	592	524	190	526	.36	1	1	1	0	0	2	0	0	0
50	1.71	1.72	523	643	624	482	51	401	1.14	1	1	0	0	0	0	1	1	1
51	2.49	2.65	690	741	704	688	41	317	1.13	1	1	0	1	0	0	0	0	1
52	1.68	1.18	569	657	634	569	80	233	1.50	1	1	0	0	0	1	1	1	1
53	2.15	1.53	543	547	530	560	8	119	- .32	1	1	0	1	0	0	0	0	0
54	1.06	1.47	675	645	559	638	136	494	.58	1	1	0	0	0	0	1	1	1
55	2.20	2.17	497	670	745	554	80	285	1.32	1	1	0	0	2	0	1	0	1
56	.91	.94	610	667	634	491	38	408	1.66	1	1	0	0	0	0	1	1	0
57	2.36	2.53	576	617	634	543	22	450	.96	1	1	0	0	1	0	0	0	0
58	1.53	1.29	628	684	662	535	45	266	1.29	1	1	0	0	0	0	1	0	1
59	2.15	1.75	515	637	606	484	50	506	.69	1	1	1	2	0	0	0	0	0
60	2.06	2.19	522	688	577	510	55	252	.78	1	1	1	1	0	0	0	1	1
61	1.81	2.00	616	647	540	553	25	474	1.58	1	1	0	0	1	0	0	1	1
62	3.53	3.59	711	737	756	676	3	350	2.39	1	1	0	0	0	1	2	1	1
63	2.21	2.31	636	679	511	428	14	299	1.68	1	1	0	0	0	0	3	1	1
64	1.31	1.27	615	651	549	517	22	164	1.11	1	1	0	0	0	0	0	0	0
65	2.06	2.12	576	737	723	754	4	144	1.92	1	1	1	1	1	2	0	0	0
66	2.03	2.56	563	677	568	491	40	185	.79	1	1	0	1	0	0	0	0	0
67	2.32	2.20	661	565	511	572	1	133	2.43	1	1	0	0	0	0	1	0	0
68	1.82	1.71	642	606	634	558	5	36	2.21	1	1	0	0	0	0	1	0	1
69	1.82	1.65	503	626	606	576	48	321	1.04	1	1	0	0	1	0	0	1	0
70	3.31	3.00	711	707	662	710	5	213	1.99	1	1	1	0	0	0	1	0	0

APPENDIX 1

TABLE 1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Business																		
71	.90	1.00	543	547	511	462	66	264	.68	1	1	0	0	1	0	0	0	1
72	2.94	3.38	563	757	745	622	93	602	1.01	1	1	0	0	1	1	0	0	1
73	.65	1.06	464	640	521	513	102	165	- .29	1	1	0	0	0	0	0	1	1
74	1.50	1.20	547	529	443	443	36	91	.27	1	1	0	0	0	0	0	1	1
75	1.60	1.40	457	616	441	539	51	148	.41	1	1	0	0	0	0	1	0	1
76	3.09	3.00	536	599	622	618	93	602	1.02	1	1	0	0	0	0	1	0	1
77	2.70	3.00	497	626	568	574	79	494	1.00	1	1	0	0	0	0	1	1	1
78	2.88	2.88	582	661	540	515	85	526	.99	1	1	0	1	0	1	0	1	1
79	2.75	2.81	569	696	524	665	29	242	1.18	1	1	1	1	0	0	0	0	1
80	2.80	2.80	484	674	536	665	53	345	1.02	1	1	0	0	0	0	1	1	1
81	2.09	2.13	477	577	549	542	31	354	1.36	1	1	0	0	0	0	1	0	0
82	2.26	2.50	615	687	615	684	34	76	.15	1	1	0	0	0	0	1	0	1
83	1.10	1.20	622	696	662	500	47	63	- .63	1	1	0	0	0	1	1	1	1
84	2.66	3.20	549	757	662	483	48	281	.95	1	1	0	0	1	0	0	1	1
Arts-Engineers																		
85	1.97	1.88	610	747	549	716	62	219	.58	1	1	0	1	0	1	0	0	1
86	1.15	1.38	603	617	549	571	41	223	.90	1	1	0	0	1	0	1	1	1
87	1.97	1.94	609	635	596	639	28	242	1.20	1	1	0	0	0	0	0	0	1
88	1.83	.88	510	556	464	440	22	291	1.44	1	1	0	1	0	0	0	0	0
89	1.58	2.00	589	625	500	608	9	69	.41	1	1	0	0	2	0	0	0	1
90	.85	.63	582	674	568	622	26	160	.99	1	1	0	0	0	0	1	1	0
91	1.79	1.88	444	626	472	486	27	440	1.55	1	1	0	0	0	3	1	1	1
92	1.71	1.72	602	757	643	581	170	642	.63	1	1	1	0	0	1	0	1	1

APPENDIX 1

TABLE 1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Arts																		
93	.84	1.40	523	581	587	660	54	382	1.08	1	1	0	0	1	0	0	1	1
94	3.21	3.19	603	727	666	463	1	30	1.85	1	1	1	0	0	0	2	0	1
95	2.12	2.29	643	747	712	589	45	311	1.06	1	1	0	0	0	2	0	1	1
96	1.36	1.59	642	739	643	587	9	35	.67	1	1	0	0	0	1	1	0	1
97	1.15	.94	642	626	559	500	19	93	.83	1	1	1	0	0	1	1	0	1
98	.42	.24	637	637	617	632	486	778	.60	1	1	0	0	1	0	0	1	1
99	3.72	3.81	691	521	535	629	134	805	.97	1	1	0	0	0	0	1	1	0
100	2.88	2.94	655	617	623	570	18	189	1.31	1	1	1	0	0	0	1	0	0

APPENDIX 1

TABLE 3

Coefficients of the Prediction Equations Developed in 1961

Curriculum	SAT Math Score	SAT Ver. Score	Secondary School Rank*	Constant
Engineering	.0037	.0043	.070	-7.170
Business	.0050	.0023	.068	-6.142
Arts	.0016	.0035	.077	-5.562
Arts-Engineering	.0073	.0011	.054	-6.788

* The secondary school rank is changed to a percentile and then normalized to a T score with mean 50 and standard deviation 10.

APPENDIX 1

TABLE 4 - Results of Regression 15 and 1961 Equations

Student Number Engineers	Achieved Cum	Predicted Cum Reg 15	Predicted Cum 1961 Equ.	Residual from Reg 15	Residual from 1961 Equ
1	2.03	2.18	2.51	- .15	- .48
2	2.91	2.82	2.75	.09	.16
3	2.64	2.48	2.86	.16	- .22
4	2.41	2.94	3.32	- .53	- .91
5	2.06	1.65	1.83	.41	.23
6	.82	1.54	1.84	- .72	-1.02
7	2.79	2.34	2.02	.45	.77
8	2.51	2.50	2.67	.01	- .16
9	1.33	1.50	1.41	- .17	- .08
10	.76	1.59	1.72	- .83	- .96
11	2.33	3.02	2.74	- .69	- .41
12	2.06	1.84	2.12	.22	- .06
13	2.79	2.02	1.78	.77	1.01
14	2.24	1.89	1.76	.35	.48
15	1.97	2.20	2.32	- .23	- .35
16	.60	.97	1.20	- .37	- .60
17	1.94	1.62	2.17	.32	- .23
18	2.41	2.67	2.04	.34	.37
19	2.36	1.72	2.31	.64	.05
20	2.42	2.48	2.41	- .06	.01
21	.25	1.22	1.44	- .97	-1.19
22	.94	1.16	1.75	- .22	- .81
23	2.15	2.01	1.75	.14	.40
24	1.00	1.41	1.93	- .41	- .93
25	1.09	1.32	1.68	- .23	- .59
26	1.57	1.65	1.86	- .08	- .29
27	1.31	2.06	2.29	- .75	- .98
28	1.67	2.43	1.77	- .76	- .10
29	2.56	1.53	1.63	1.03	.93
30	1.21	1.52	1.54	- .31	- .33
31	2.79	2.41	2.48	.38	.31
32	1.76	1.57	1.91	.19	- .15
33	3.31	2.59	2.98	.72	.33
34	3.19	2.29	2.35	.90	.84
35	.67	.95	1.23	- .28	- .56

APPENDIX 1

TABLE 4 - Results of Regression 15 and 1961 Equations

Student Number Engineers	Achieved Cum (con't)	Predicted Cum Reg 15	Predicted Cum 1961 Equ.	Residual from Reg 15	Residual from 1961 Equ
36	.75	1.05	1.49	- .30	- .74
37	1.12	1.30	1.61	- .18	- .49
38	1.17	2.39	2.16	-1.22	- .99
39	2.03	2.26	2.04	- .23	- .01
40	2.21	2.18	2.05	.03	.16
41	1.18	2.38	2.18	-1.20	-1.00
42	1.28	.86	1.11	.42	.17
43	1.15	1.62	1.79	- .47	- .64
44	2.32	2.19	3.00	.13	- .68
45	2.68	1.70	2.09	.98	.59
46	.60	1.89	1.95	-1.29	-1.35
47	3.91	2.71	3.10	1.20	.81
48	1.15	1.72	1.79	- .57	- .64
49	.53	.81	1.24	- .28	- .71
50	1.71	1.98	1.75	- .27	- .04
51	2.49	2.50	2.82	- .01	- .33
52	1.68	1.24	1.49	.44	.19
53	2.15	1.38	1.72	.77	.43
54	1.06	2.04	2.04	- .98	- .98
55	2.20	1.57	1.35	.63	.85
56	.91	2.05	2.34	-1.14	-1.43
57	2.36	2.07	2.23	.29	.13
58	1.53	2.02	2.23	- .49	- .70
59	2.15	1.87	1.79	.28	.36
60	2.06	1.80	1.66	.26	.40
61	1.81	2.49	2.49	- .68	- .68
62	3.53	3.30	3.73	.23	- .20
63	2.21	2.57	2.73	- .36	- .52
64	1.31	1.66	2.16	- .35	- .85
65	2.06	2.29	2.84	- .23	- .78
66	2.03	1.49	1.81	.54	.22
67	2.32	3.13	2.91	- .81	- .59
68	1.82	1.96	2.10	- .14	- .28
69	1.82	1.50	1.53	.32	.29
70	3.31	2.72	3.36	.59	- .05

APPENDIX 1

TABLE 4 - Results of Regression 15 and 1961 Equations

Student Number	Achieved Cum	Predicted Cum Reg 15	Predicted Cum 1961 Equ.	Residual from Reg 15	Residual from 1961 Equ
Business					
71	.90	1.38	1.70	- .48	- .80
72	2.94	2.69	3.03	.25	- .09
73	.65	.91	1.33	- .26	- .68
74	1.50	1.03	1.35	.47	.15
75	1.60	1.33	1.67	.27	- .07
76	3.09	2.10	2.17	.99	.92
77	2.70	2.10	2.20	.60	.50
78	2.88	2.19	2.57	.69	.31
79	2.75	2.45	2.84	.30	- .09
80	2.80	2.23	2.43	.57	.37
81	2.09	1.79	2.15	.30	- .06
82	2.26	1.53	2.21	.73	.05
83	1.10	.81	1.74	.29	- .64
84	2.66	2.56	2.95	.10	- .29
Arts-Engineers					
85	1.97	1.54	2.35	.43	- .38
86	1.15	1.36	1.56	- .21	- .41
87	1.97	1.68	1.86	.29	.11
88	1.83	1.20	1.30	.63	.53
89	1.58	1.45	1.73	.13	- .15
90	.85	1.36	2.00	- .51	-1.15
91	1.79	1.43	1.79	.36	- .00
92	1.71	1.91	2.44	- .20	- .73
Arts					
93	.84	1.72	1.87	- .88	-1.03
94	3.21	2.80	2.95	.41	.26
95	2.12	2.21	2.54	- .09	- .42
96	1.36	1.84	2.25	- .48	- .88
97	1.15	1.64	2.18	- .49	-1.03
98	.42	1.26	1.29	- .84	- .87
99	3.72	3.09	2.28	.63	1.44
100	2.88	2.14	2.57	.74	.31

APPENDIX 1 - TABLE 5

Correlation Coefficients

Variables	SATM	SATV	DIF	AVEACH	CON HS	FL?	AID?	#OB	#OS	#YB	#YS	MCOL?	FCOL?	FCUM	FCUM - COLAVE	HSIZE
SATM																
SATV	27															
DIF	46	-55														
AVEACH	61	36	19													
CON HS	5	3	6	26												
FL?	- 9	5	-11	1	-12											
AID?	23	15	8	16	8	-10										
#OB	- 9	0	1	1	8	6	31									
#OS	- 4	-13	2	7	- 3	7	-14	0								
#YB	16	-11	19	- 4	- 9	-10	17	0	-20							
#YS	13	27	- 7	11	9	0	- 4	-18	-21	3						
MCOL?	6	- 4	3	- 7	-17	- 3	-16	-21	9	0	21					
FCOL?	20	- 1	9	- 4	-29	-12	-17	-21	10	5	20	37				
FCUM	32	25	16	44	49	- 8	9	5	- 8	-17	20	5	5			
FCUM-COLAVE	34	27	15	45	53	- 9	10	6	- 8	-15	21	4	2	-		
HSIZE	- 3	- 4	8	10	16	2	- 7	- 5	- 7	7	-12	9	- 3	17	16	
HPOS	- 5	7	- 8	- 1	-56	7	- 6	- 9	7	5	- 8	19	17	-21	-24	57
#SIBS	22	3	10	-	2	- 2	23	-	20	62	45	7	10	0	-	- 9
ENG	8	- 2	- 4	-	38	- 9	7	10	- 5	7	20	-15	32	4	-	- 7

HPOS-ENG = 26
HPOS-#SIBS = - 4

Figures in table are the correlation coefficients multiplied by 100 and rounded off.

APPENDIX 1

TABLE 6 - STATISTICAL SUMMARY OF VARIABLES

Variable	Range of Values	Mean	Variance	Standard Dev.
SATM	444-757	657.6	3992	63.2
SATV	495-800	567.5	4298	65.6
DIF	0-225	100.4	4050	63.6
AVEACH	443-764.5	585.9	4326	65.8
CON HS	-.63-2.43	1.1369	.292	.54
FL?	0,1	.98	.020	.14
AID?	0,1	.21	.17	.41
#OB	0,1	.21	.27	.52
#OS	0,1	.23	.24	.49
#YB	0,1	.48	.51	.72
#YS	0,1	.47	.41	.64
MCOL?	0,1	.39	.24	.49
FCOL?	0,1	.59	.24	.49
FCUM(S'62)	.25-	1.90	.655	.81
FCUM-COLAVE	-1.62-2.04	.0057	.64	.80
HSIZE	19-805	285.7	34551	185.8
HPOS	1-190	43.5	3882	58.1
#SIBS	0-6	1.39	1.07	1.03
ENG	0,1	.70	.21	.46
DIF-50	0,200	71.2	2533	50.3
MAXACH	443-800	572.8	4050	63.6

APPENDIX 2

REGRESSION NO. 3

Dependent Variable = FCUM

R**2 (Maximum likelihood) .44318358 00
R**2 (Based on unbiased variances) .40081712 00
Residual Mean Square = .39248383 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.70072771 00	.12764454 00	.30136588 02	3
AVEACH	.16323056-02	.13324336-02	.15007619 01	0
FCOL?	.25309154 00	.13738515 00	.33937174 01	1
HSIZE	.76497871-03	.34587326-03	.48917554 01	2
SATV	.15981104-02	.10415089-02	.23544398 01	0
#YB	-.17907377 00	.91618695-01	.38202849 01	1
SATM	.23526378-02	.13534828-02	.30213749 01	1

Constant term = -.25898908 01

REGRESSION NO. 4

Dependent Variable = FCUM

R**2 (Maximum likelihood) .44961935 00
R**2 (Based on unbiased variances) .40774256 00
Residual Mean Square = .38794744 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.67605037 00	.12698376 00	.28344039 02	3
AVEACH	.24809027-02	.10995883-02	.50904859 01	2
FCOL?	.30396752 00	.13361658 00	.51752780 01	2
HSIZE	.70304742-03	.34103255-03	.42498888 01	2
SATV	.27611957-02	.11377911-02	.58893765 01	2
DIF*	.28255915-02	.13899639-02	.41324887 01	2
#YB	-.14985427 00	.88450837-01	.28703467 01	1

Constant term =-.23997643 01

*DIF = Math - Verb - 50

APPENDIX 2

REGRESSION NO. 6

Dependent Variable = FCUM

R**2 (Maximum likelihood) .47619701 00
R**2 (Based on unbiased variances) .43634243 00
Residual Mean Square = .36921362 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.68811018 00	.12322457 00	.31183247 02	3
MAXACH	.10860920-02	.90007870-03	.14560365 01	0
FCOL?	.28011029 00	.13040868 00	.46136548 01	2
SATV	.48047689-02	.11962046-02	.16133704 02	3
DIF	.43924102-02	.12180060-02	.13004897 02	3
#YB	-.18652653 00	.87370088-01	.45578045 01	2
HSIZE	.67231556-03	.33314234-03	.40727399 01	2

Constant Term = -.29411606 01

REGRESSION NO. 7

Dependent Variable = FCUM

R**2 (Maximum likelihood) .47110618 00
R**2 (Based on unbiased variances) .43086426 00
Residual Mean Square = .37280199 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.71216977 01	.12122595 00	.34512455 02	3
SATV	.51878773-02	.11397148-02	.20719871 02	3
DIF	.46791714-02	.11942891-02	.15350374 02	2
#YB	-.24495974 00	.11301237 00	.46982644 01	2
HSIZE	.76376448-03	.33660463-03	.51484766 01	2
FCOL?	.27098286 00	.13185823 00	.42234678 01	2
#SIBS*	.58201469-01	.78020374-01	.55648290 00	0

Constant Term = -.26661845 01

* Number of Siblings

APPENDIX 2

REGRESSION NO. 8

Dependent Variable = FCUM

R**2 (Maximum likelihood) .44409659 00
R**2 (Based on unbiased variances) .40823185 00
Residual Mean Square = .38762694 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.75280916 00	.12212543 00	.37997720 02	3
SATV	.53751634-02	.11588104-02	.21515854 02	3
DIF	.44365535-02	.12124435-02	.13389625 02	3
FCOL?	.29059601 00	.13413747 00	.46933083 01	2
HSIZE	.65796988-03	.33960463-03	.37537454 01	1
#SIBS*	-.48487152-01	.61726872-01	.61702793 00	0

Constant term = -.27449280 01

* Number of Siblings

REGRESSION NO. 11

Dependent Variable = FCUM

R**2 (Maximum likelihood) .44127506 00
R**2 (Based on unbiased variances) .39876339 00
Residual Mean Square = .39382910 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.77898416 00	.12523357 00	.38691514 02	3
SATM	.41743629-02	.11626968-02	.12889875 02	3
CURR*	.25836807 00	.11978409 00	.46524258 01	2
HSIZE	.83938577-03	.34241963-03	.60090511 01	2
#YB	-.18609966 00	.91919584-01	.40989730 01	2
FCOL?	.21977300 00	.13734771 00	.25603883 01	0
DIF	-.55661614-03	.11577075-02	.23116065 00	0

Constant Term = -.19506767 01

* CURR = Curriculum: Eng=0; Bus=+1; A-AE=-1

APPENDIX 2
REGRESSION NO. 12

Dependent Variable = FCUM

R**2 (Maximum Likelihood) .48131059 00
R**2 (Based on unbiased variances) .44184509 00
Residual Mean Square = .36560921 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.79593577 00	.12048203 00	.43623688 02	3
SATM	.30219707-02	.10602405-02	.81240338 01	3
CURR*	.34271054 00	.11843223 00	.83736562 01	3
SATV	.27791878-02	.10251113-02	.73501088 01	3
HSIZE	.84341983-03	.32827194-03	.66011574 01	2
FCOL?	.24255406 00	.13260549 00	.33457544 01	1
#YB	-.14312863 00	.89385392-01	.25640114 01	0

Constant Term = -.28788101 01

*CURR = Curriculum: Eng=0; Bus=+1; A-AE=-1

APPENDIX 2
REGRESSION NO. 13

Dependent Variable = FCUM

R**2 (Maximum likelihood) .53474426 00
R**2 (Based on unbiased variances) .45166288 00
Residual Mean Square = .35917825 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.65438263 00	.17103887 00	.14637735 02	3
SATM	-.21470576-03	.16373407-02	.17195281-01	0
CURR*	.33076500 00	.12237811 00	.73051912 01	3
SATV	.57987880-02	.17294346-02	.11242585 02	3
DIF	.44871820-02	.18173728-02	.60962018 01	2
HSIZE	.10925671-02	.48902488-03	.49915372 01	2
FCOL?	.25352712 00	.14850359 00	.29145728 01	1
#YB	-.16248481 00	.92898314-01	.30592139 01	1
#YS	.10238290 00	.10851664 00	.89014788 00	0
HPOS	-.17026615-02	.18698317-02	.82918528 00	0
#OB	.70456904-01	.12858834 00	.30022296 00	0
AID?	.79588993-01	.16836676 00	.22345666 00	0
MCOL?	.62726120-01	.13848579 00	.20515705 00	0
FL?	-.16668579-01	.44687285 00	.13913270-02	0
#OS	-.18602360-02	.13658903 00	.18548314-03	0

Constant Term = -.28356770 01

*CURR = Curriculum: Eng=0; Bus=+1; A-AE=-1

APPENDIX 2

REGRESSION NO. 14

Dependent Variable = FCUM

R**2 (Maximum likelihood) .41424286 00
R**2 (Based on unbiased variances) .38957940 00
Residual Mean Square = .39984490 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.75842268 00	.11905720 00	.40579931 02	3
SATM	.30631199-02	.10486390-02	.85324939 01	3
CURR*	.38088649 00	.12137658 00	.98443939 01	3
SATV	.29134904-02	.10464093-02	.77521831 01	3

Constant Term = -.26232476 01

*CURR = Curriculum: Eng=0; Bus=+1; A-AE=-1

REGRESSION NO. 16

Dependent Variable = FCUM-Mean College Cum.

R**2 (Maximum likelihood) .51453259 00
R**2 (Based on unbiased variances) .47759485 00
Residual Mean Square = .34924091 00
Degrees of Freedom = 88

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.75045598 00	.11764634 00	.40690553 02	3
AVEACH	.12956711-02	.11234726-02	.13300395 01	0
FCOL?	.24609488 00	.12441972 00	.39122508 01	1
SATV	.45843211-02	.12809225-02	.12808678 02	3
DIF	.38465857-02	.12781516-02	.90570359 01	3
HSIZE	.64558430-03	.31637889-03	.41638103 01	2
#YB	-.14996641 00	.83198195-01	.32490781 01	1

Constant Term = -.48523233 01

APPENDIX 2

REGRESSION NO. 17

Dependent Variable = FCUM-Mean College Cum.

R**2 (Maximum likelihood) .46802345 00
R**2(Based on unbiased variances) .42754697 00
Residual Mean Square = .38269936 00
Degrees of Freedom = 88

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.74677263 00	.12343112 00	.36603909 02	3
AVEACH	.19746014-02	.12679431-02	.24252658 01	0
FCOL?	.20943867 00	.13238130 00	.25029978 01	0
HSIZE	.69837289-03	.33609500 00	.43176820 01	2
SATM	.27518737-02	.14412724-02	.36455649 01	1
#YB	-.16607652 00	.88275854-01	.35394244 01	1
DIF	-.10858562-03	.11066641-02	.96274878-02	0

Constant Term = -.40423151 01

REGRESSION NO. 18

Dependent Variable = FCUM-Mean College Cum.

R**2 (Maximum likelihood) .48709667 00
R**2 (Based on unbiased variances) .44807142 00
Residual Mean Square = .36879794 00
Degrees of Freedom = 89

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.76370964 00	.12104285 00	.39808725 02	3
AVEACH	.14550508-02	.12635209-02	.13261462 01	0
FCOL?	.22560406 00	.13027968 00	.29987514 01	1
SATV	.18294518-02	.98764271-03	.34311697 01	1
HSIZE	.73359997-03	.32798491-03	.50027738 01	2
SATM	.24525429-02	.12834816-02	.36513540 01	1
#YB	-.14749095 00	.86880235-01	.28819654 01	1

Constant Term = -.46380068 01

APPENDIX 2
REGRESSION NO. 19

Dependent Variable = FCUM-Mean College Cum.

R**2 (Maximum likelihood) .54088437 00
R**2 (Based on unbiased variances) .45889944 00
Residual Mean Square = .36331217 00
Degrees of Freedom = 80

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.55556995 00	.16733551 00	.11023036 02	3
AVEACH	.18847239-02	.13022498-02	.20946279 01	0
FCOL?	.22699226 00	.14573052 00	.24261690 01	0
SATV	.42918305-02	.16812158-02	.65168572 01	2
DIF	.36267596-02	.17822414-02	.41409928 01	2
HSIZE	.11460333-02	.47600209-03	.57966455 01	2
#YB	-.16727116 00	.89193203-01	.35170506 01	1
HPOS	-.26977050-02	.18032393-02	.22381138 01	0
#YS	.97018249-01	.10631767 00	.83271412 00	0
#OB	.11063652 00	.12672902 00	.76215778 00	0
MCOL?	.98675094-01	.13761335 00	.51415484 00	0
#OS	-.53783118-01	.13507822 00	.15853359 00	0
FL?	-.11818272 00	.44022648 00	.72070188-01	0
SATM	-.47546240-03	.17483818-02	.73953680-01	0
AID?	.26388702-01	.16506729 00	.25557249-01	0

Constant Term = -.44658361 01

APPENDIX 2

REGRESSION NO. 21*

Dependent Variable = FCUM-Mean College Cum.

R**2 (Maximum likelihood) .60155030 00
R**2 (Based on unbiased variances) .54595267 00
Residual Mean Square = .32218116 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.76028028 00	.17625224 00	.18607081 02	3
HSIZE	.16566567-02	.45601332-03	.13198056 02	3
SATV	.52542146-02	.12538524-02	.17559930 02	3
DIF	.42664241-02	.14774456-02	.83388289 01	3
#YB	-.69582612-01	.11479620 00	.36740618 00	0
FCOL?	.10421700 00	.17340954 00	.36118622 00	0

Constant Term = -.47896237 01

* 50 Data points Group 1

REGRESSION NO. 22*

Dependent Variable = FCUM-Mean College Cum.

R**2 (Maximum likelihood) .49637097 00
R**2 (Based on unbiased variances) .42509716 00
Residual Mean Square = .33327835 00

Variable	Coefficient	S. D. Coeff.	F(T**2)	Conf. LVL
CON HS	.75860949 00	.15787395 00	.23089265 02	3
FCOL?	.29022607 00	.19937488 00	.21190050 01	0
#YB	-.20053944 00	.12135825 00	.27306184 01	0
SATV	.48477983-02	.21947436-02	.48788956 01	2
DIF	.43187138-02	.19825736-02	.47451528 01	2
HSIZE	-.88984695-04	.46115769-03	.37233317-01	0

Constant Term = -.40748432 01

* 50 Data points Group 2

INTERVIEWS

Dr. Arthur L. Brody; Psychology Department, Lehigh University

Mr. Forrest Jewell; Psychology Department, Carnegie Institute of Technology

Mr. Samual H. Missimer; Director of Admission, Lehigh University

Dr. Melvin Novick; Educational Testing Center, Princeton, New Jersey

BIBLIOGRAPHY

Bradley, Gordon H., LEHIGH UNIVERSITY UNDERGRADUATE INFORMATION SYSTEM, Thesis, Lehigh University, 1964

College Entrance Examination Board, A DESCRIPTION OF THE 1965-66 COLLEGE BOARD VALIDITY STUDY SERVICE, a booklet prepared by the College Entrance Examination Board, 1965

Duggan, J. M., Hazlett, P. H., Jr., PREDICTING COLLEGE GRADES, A COMPUTATION WORKBOOK FOR ESTIMATING FRESHMAN GRADE AVERAGES FROM HIGH SCHOOL RECORDS AND COLLEGE BOARD SCORES, College Entrance Examination Board, New York, 1961

Ezekiel, M., Fox, K. A., METHODS OF CORRELATION AND REGRESSION ANALYSIS LINEAR AND CURVILINEAR, John Wiley and Sons, Inc., New York, 1959, Third Edition

Ferguson, G. A., STATISTICAL ANALYSIS IN PSYCHOLOGY AND EDUCATION, McGraw-Hill Book Company, Inc., New York 1959

Gulliksen, H., THEORY OF MENTAL TESTS, John Wiley and Sons, Inc., New York, 1950

Polya, G., "Remarks on Computing the Probability Intergral in One and Two Dimensions", PROCEEDINGS OF THE BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, University of California Press, Berkeley and Los Angeles, 1949

Ponhoff, R. F., THE PREDICTION OF COLLEGE GRADES FROM COLLEGE BOARD SCORES AND HIGH SCHOOL GRADES, (University of North Carolina) Institute of Statistics, Mimeo Series #419, Department of Statistics, University of North Carolina, Chapel Hill, North Carolina, 1964

Quenouille, M. H., INTRODUCTORY STATISTICS, Pergamon Press Ltd., London, 1950

Shepard, R. N., "Extracting Latent Structure from Behavioral Data", PROCEEDINGS OF THE 1964 SYMPOSIUM ON DIGITAL COMPUTING, Holmdel Laboratory, Bell Telephone Laboratories, January 30-31, 1964

Wert, J. E., Neidt, C. O., Ahmann, J. S., STATISTICAL METHODS IN EDUCATIONAL AND PSYCHOLOGICAL RESEARCH, Appleton-Century-Crofts, Inc., New York, 1954

Wilson, E. B., AN INTRODUCTION TO SCIENTIFIC RESEARCH, McGraw-Hill Book Company, Inc., New York, 1952.

VITA

David Anthony Riemondy born May 2, 1942 in Reading, Pennsylvania to Mary E. Riemondy and Col. Augustus A. Riemondy, USAF; received a Bachelor of Science in Industrial Engineering from Lehigh University in June, 1964; part-time and summer industrial engineering employment with Federal Civil Service, July and August of 1962 and 1963; the Taylor-Wharton Company, March 1964 through June, 1965; the Bethlehem Steel Corporation, June to September, 1965; full time freshman counselor in the Residence Halls at Lehigh University for six semesters from September, 1961 to June, 1964; part-time teaching assistantship in the Industrial Engineering Department of Lehigh University for three semesters from September, 1964 to January, 1966; commissioned as a second lieutenant in the United States Air Force, June, 1964; member of Alpha Phi Mu.